



中文使用手册

Cao Wei (caowei_888@sina.com) 等 翻译
生物软件网 (<http://www.bio-soft.net>)

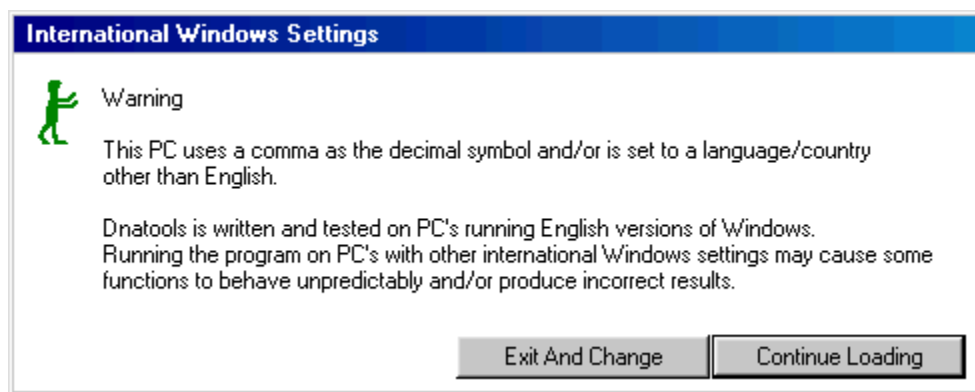
整理提供

Chapter1: about DNAtools

1. 视窗设置，国家，语言：

十进制分割问题，微软：

我最近发现的 DNAtools 中的一个小 Bug (Microsoft Visual Basic 处理国家特异的十进制分割符号的方式)。使用 “*Windows/Control Panel*” 中除了英语之外的国家设置将可能导致 Visual Basic 错误的解释十进制值。为了暂时回避这些问题，我已经进行了针对国家/语言和激活的十进制分离子的测试，且在装载 DNAtools 的第一张表后马上进行该项测试。除非 PC 设置位英语（任何版本）并且使用完全的终止作为十进制分离器，否则用户将会看到以下这个警告（当用户启动 DNAtools 时）。



如果看到这个，用户可以：

忽略它并继续使用；

或者在 “*Windows/Control*” 中改变设置。

然而，记住与十进制值有关的功能将产生不正确的结果或者奇怪的执行命令（如果用户决定使用非英语的国家设置）

2. 关于 DNAtools 帮助：

DNAtools 不提供打印的手册。因为整个 DNAtools 组织只包含一个人，不太可能维持这些代码、帮助文本和个人主页。

尽管我尽力将帮助文本整合入程序中，升级总是落后几个版本。

如果用户希望获得关于程序工具的总结，查看帮助内容，在线手册和主页中的“New in This version”部分。若还是不能找到说资料，可写 E-mail 给我，我很乐意帮助你。

程序的附件，补丁和帮助文档的提取信息可在主页上找到，有时可在 bionet.software 新闻组中找到。

Soeren W. Rasmussen

February 9th, 2000

3. 关于序列名字：

DNATools 使用旧 DOS 文件名。这里解释为什么这样做。

为了全面的利用几个功能（这些功能是为了在一个方案中处理多个文件时而设计的），有必要理解 DNATools 是如何在这些功能中使用文件名的。这个帮助文档尽力解释为什么在命名序列时相容的行为是这些功能正确的恰当的工作的前提条件。

背景：

在 DNATools 中大多数的与处理多个序列相关的功能是在一个小的执行于 Carlsberg 实验室的 EST 项目中发展而来的。此项目旨在获得关于 Blumeria（霉菌）的基因组信息和基因的表达情况以更好的理解植物寄生物和其宿主大麦之间的交互作用。

用前向和反向引物，对所有的来自 cDNA 库并用于分析的克隆测序两次。插入的 5' 序列用于在公开的数据库中进行数据库同源性查找。而 3' polyA 序列则用于生成连接到产生于同样发展阶段的 SAGE。

CDNA 库的插入长度是非常短的（只有很少的是全长的 ORFs），这反而成为一个优点，尤其是用于寻找国际性的数据库时。它同时也暗示了在许多场合下，F 和 R 序列是交叠的且可以被这个特殊插入其结合的完全的序列所代替。库的特征允许我们用

其结合的序列替换 F 和 R 序列，这些结合的序列既提高了序列的质量也同时降低了 Blumeria 数据库的序列数据量。

万一某个克隆/插入的 F 和 R 序列并没有交叠，例如用于连接来自同一克隆的 F 和 R 序列的序列信息是不可获得的，则使用文件的名称来代替同一插入的 F 和 R 序列的连接。很明显的，这个需要文件/克隆必须始终按照下面描述的进行命名。

文件名字：

为了跟踪起始于同一插入/插入的 F 和 R 序列，所有的序列需要依据 DOS 结构命名原则进行命名，在这种情况下，一个文件名包含一个八字符名，一点和三字符扩展名（例如 NAMENAME.EXT）

名字前六个字符表示克隆或插入 (e. g. C00018, ABCDEF, 012345).

第七个字符对于 DNAtools 来说不被考虑，它可以是一个破折号来填满字符数，或者被用于含同样 F 或 R 引物的多重序列命名，或者用于引物步行的二级引物命名 (e. g. C00018-, ABCDEF-, 012345-, C00018a, ABCDEFa, 012345a)

第八个字符用于确定起始的是 F（前向，沃森，上游链）还是 R（反向，克里克，下游链）（e. g. C00018-F, C00018-R, ABCDEFaF, ABCDEFaR, 012345bF, 012345bR）.

自动的或者是用户整合的 F 和 R 序列接受可以被 DNAtools 识别的特征 M (e. g. C00018-M, ABCDEFaM, 012345bM)。除了 F, R 和 M 之外，在名字的第八个字符是不被 DNAtools 识别的。

三字符的扩展名作为特殊的参数是不被 DNAtools 识别的，但它可以被用于增加额外的信息到文件名中且并不影响多重序列操作。（e. g. C00018-F.seq, C00018-R.old, ABCDEFaF.new）

DNAtools DOS file name: **CCCCC X S . XXX**

综上：

红色参数对于 DNAtools 有专门的意义（C=克隆名字，S=链），而绿色参数（X=选项）作为特殊的参数是不被 DNAtools 识别的。

长序列名字：

装载一个新的长名字的非 DOS 文件名的序列到 DNAtools 中，DNAtools 会自动的将长文件名传输到 DNAtools 的长文件名变量中。接着长文件名被加工后以产生 DOS 名。最后 DOS 名被传输到 DNAtools 的名字变量中。

对于那些含正确恰当的 DOS 文件名字的新的序列，而这些序列又没有事先被 DNAtools 格式化，一个长文件名会自动的被生成。且这个文件名包含“#”键，后跟 8 个数字。（e. g. C00018-F #47382957）

如果用户希望改变长文件名为一个更加有描述性的序列名，可以通过在常规的标题形式中实现。

长文件名（用于 DNA 和蛋白质和引物序列）包含于所有的印刷物中且和序列文件同时保存。对于引物序列，如果用户想用 DNAtools 进行 E-mail 订购引物时，长文件名是出现于订单形式中唯一的名字。

当行展示选项中定义为行 0 时，在常规的序列标题和主要的编辑形式的信息栏中，长文件名是可见的。看下面，标题行 1-5，用于其他的方法确定展示的序列。

从 DNAtools 不太可能保存含长文件名的序列。

改变文件名：

如果文件名不符合 DNAtools 的准则或缺乏内在的逻辑时，DNAtools 有三个选项用于操纵文件名以帮助用户改变文件名，

如果希望改变已存在的文件名同时保留文件名的起始部分，使用 `change file names;`

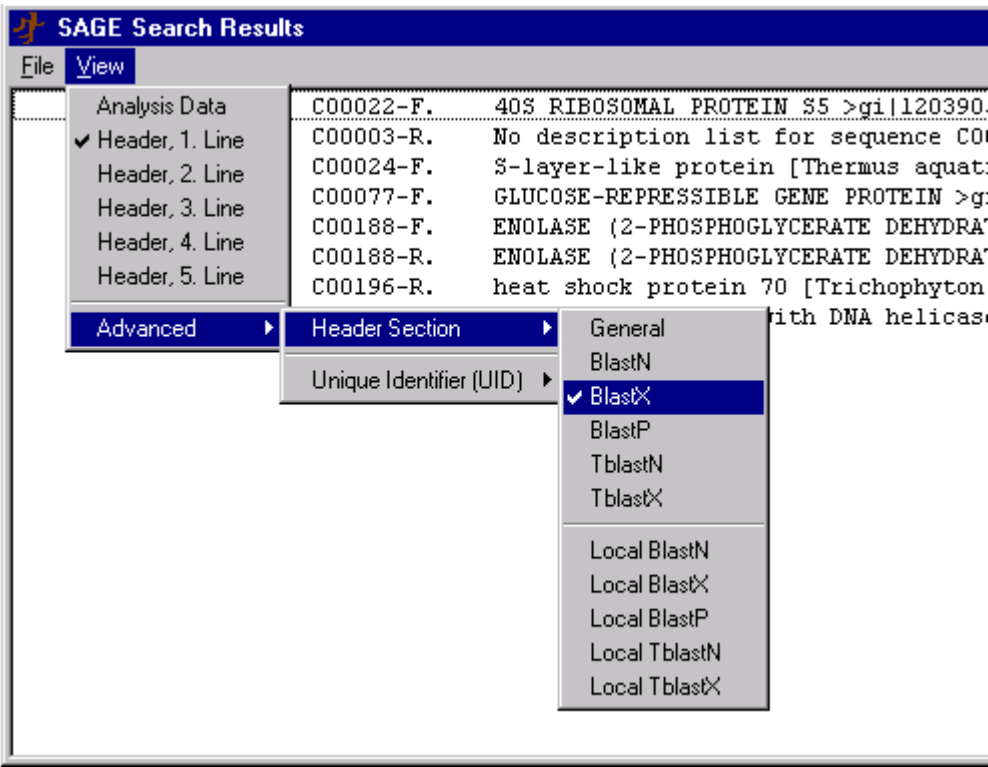
如果希望基于相同的模板生成一个完全新的文件名，使用 `Renaming template;`

如果希望独立的编辑任何一个文件名，使用 Manual editing

标题行 1—5:

很明显的，当一个方案包含几千个序列时，不太可能依据 DOS 文件名确定或重新确
记一个单个的序列。为了帮助用户记忆，DNAtools 可以使用户功能展示序列中展示
序列文本标题的一个选择的行（行 1—5 是指定的引导部分）。使用 View 选项获得菜
单结构，显示如下。

选择用户期望使用的引导行，通常行 1 包含最好的数据库匹配。接着去高级
选项，选择一个特殊的部分。最后用户可以选择是否希望包含 UID 在展示行
中。在大多数情况下，在文件列表中，对于达到“确定”目的，这种做法是
没有用的。



如果用户已经执行了（例如，一个 blastx 寻找同时在重新得到寻找结果前也已经审查
了自造的选项，将会为每个序列自动的生成一个包含 blastx 描述行的引导部分）。设
置 View 菜单选项，显示所有形式以产生序列列表，接着 Blastx 行 1，然后使形式针对

所有的序列显示最好的 blastx 数据库匹配。这些使得用户很容易鉴定独立的序列，即使该方案包含大量的序列数据。

4. 关于方案：

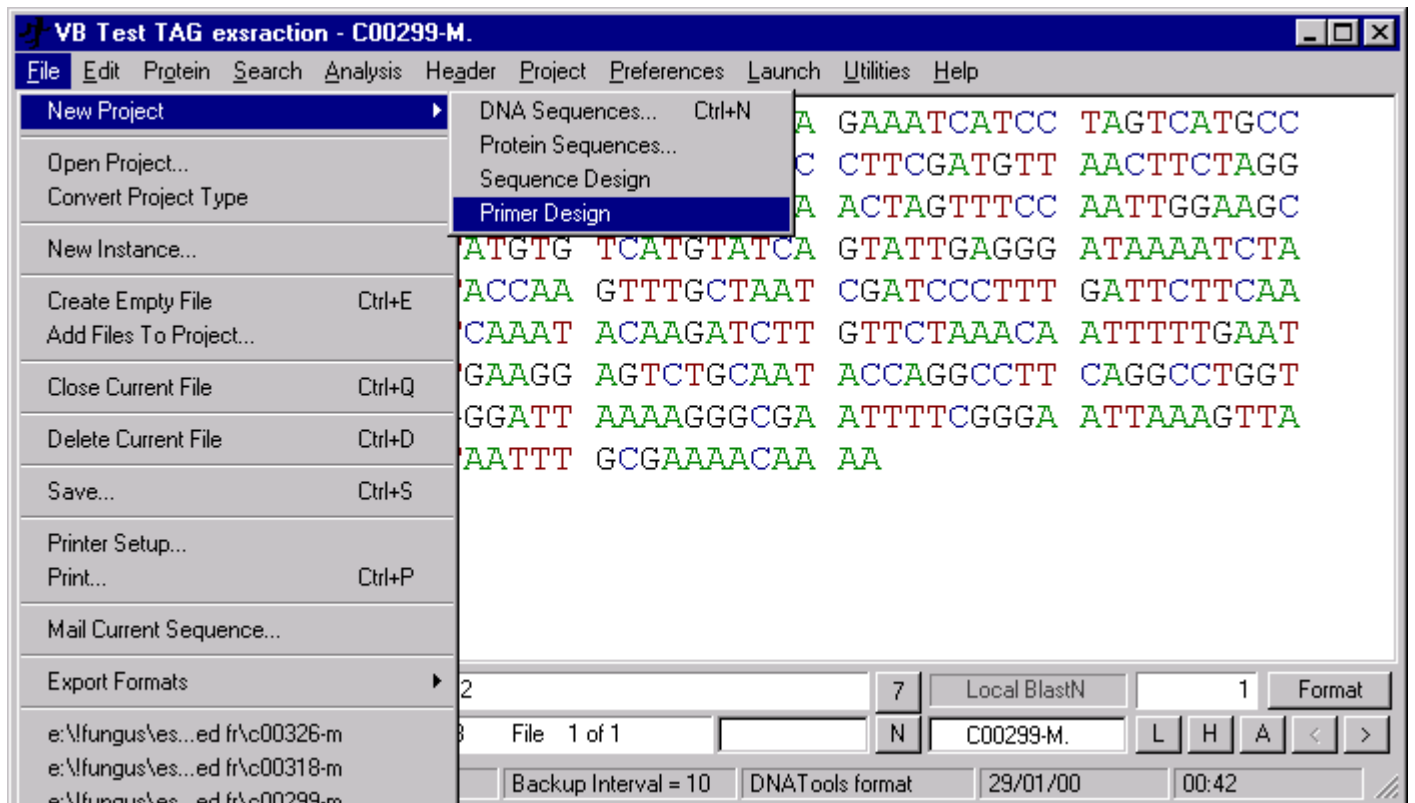
可以打开一个方案（例如一些装载于 DNAtools 中的序列）作为以下四种类型之一：

DNA 序列方案：

这种类型是用于 DNA 序列和已经存在的寡核苷酸引物文件。DNAtools 自动的鉴别正在装载的是正常的序列文件还是引物文件。如果用户尝试装载一个序列和引物文件的混合物到同一个方案中，将会提示错误。为了避免这些，装载正常的序列文件和引物序列文件到不同的方案中。注意，用户可以同时运行分离的、独立的 DNAtools 实例。万一一个引物序列被错误的保存为正常序列，它有可能被转化这种类型的文件。

观看所有的/观看 psg 按钮：

除了可以选择单个序列文件，psg 文件可以被选择并从主要的装载表中被装载。然而，从 psg 方案文件中一次只能装载一个方案。看增加文件到方案，在打开的方案中寻求更多的信息。在文件列表中的两个展示之间点击 *View psg/View all* button。



蛋白质方案：

蛋白质方案用于蛋白质文件。功能是有限的，但是包含选项用于重命名和改变文件格式。

序列设计方案：

当用户想手动进入新的序列或从其他数据源拷贝/粘贴序列时才使用这种方案。不像前面两种方案，当一个新的序列设计方案产生时，用于装载文件的格式是不被显示的。如果用户想增加已经存在序列文件到用户的序列设计方案中时，使用 *File/Add Files To Project* 选择需装载的文件。

引物设计方案：

当用户想创造新的引物时采用这种类型的方案，不像前面两种方案，当一个新的引物设计方案产生时，用于装载文件的格式是不被显示的。如果用户想增加已经存在引物文件到用户的引物设计方案中时，使用 *File/Add Files To Project* 选择需装载的文件

注意主菜单的菜单项目是随着方案类型而改变的。如果针对当前序列文件或在新的方案被打开前他们的名字发生改变，将会提示警告。一个方案的最大序列数量（即可以被装载到一个方案中的文件）的数量是有限的，依赖于内存的大小。然而，当装载的数据量接近一百万时程序速度会急剧的下降。

单个序列的最大长度是有限的，依赖于可获得内存的大小。然而，随着长度的增加，程序的速度急剧下降。一百万个数据的序列可以被重新格式化，寻找，翻译等等，当然前提是用户有足够的耐心。

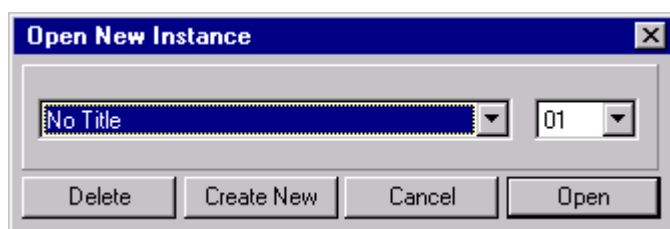
在装载前对序列文件进行确认—在新的序列文件被装载到一个方案中时，文件的内容需要被检测以求一个有效的正确的分离物分离引导序列和 DNA 序列同时也为了寻找在文件的序列部分中不合法则特征存在的可能性。

如果一个有效的正确的分离物缺失的话，将会提示警告。如果文件已经被用户所接受，默认的 DNAtools 分离物被自动的加载到序列的前面。

在文件被装载到方案中前或者文件可以被跳过，序列部分中无效的特征可以被显示和编辑。如果一个包含无效特征的文件在没有被编辑前就被装载，且当序列是被格式化的，那么这些无效的特征则被转化为 N 或者在没有进一步警告的情况下被移除。

DNAtools 额外的实例：

出于某些目的，打开一些额外的 DNAtools 程序会更好，而不是仅仅打开一个新的项目（方案）。在这种情况下，两个或更多组的序列文件可以同时被编辑，序列信息可以在不同的方案中拷贝或粘贴。DNAtools 中活动实例的数量是有限的，依赖于计算机的内存。有可能打开一个正常序列方案实例，同时又打开一个引物设计实例或者蛋白质方案实例。



5. 关于 basecalling:

当 basecalling 已存在的 ABI 格式的色谱图或者从一系列格式的色谱图中提取纯 ASCII 序列（无重复 basecalling）时，此表允许用户控制使用的参数。

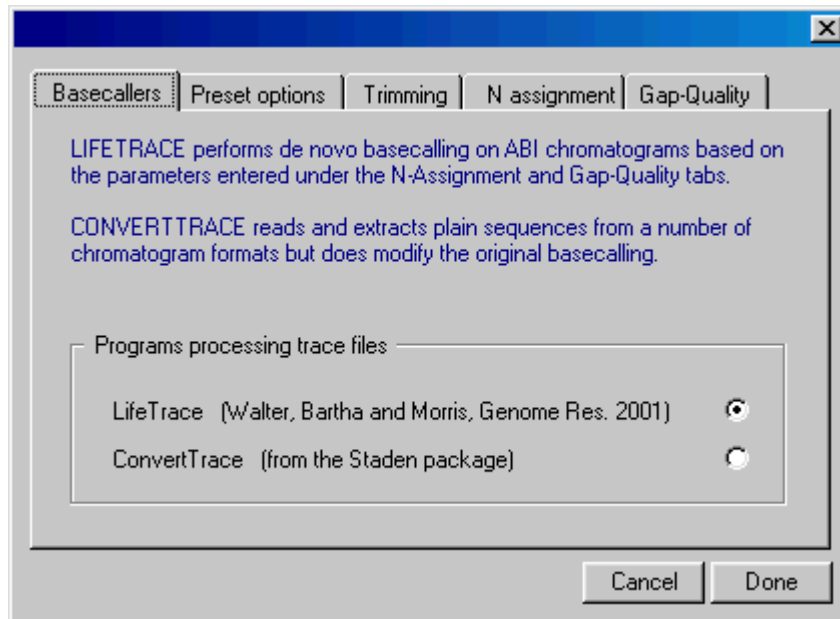
可以查看以下网页：

- [LifeTrace - Basecalling Software](#)
- [LifeTrace on Linux/Unix systems](#)
- [View trace files with Chromas](#)

卓越的 basecaller—lifeTrace (Dirk Walters 等人出版 in Genome Research, 11, 875-888, 2001) 将很快可以获得一个 Win32 版本。DNAtools 的界面已经被写好并被测试了，但是在 Win32 版本可以公开获得之前，版权问题仍未解决。相比于 Phred，基准测试证明 lifetrace 可以产生小于 17% 置换错误，小于 16% 的插入/删除错误。另外一个优点，lifetrace 产生所谓的 Gap_q 分数，提供邻近碱基之间 Gap 的质量评价。所有的 Lifetrace 特征都可以通过 DNAtools 使用者界面获得，见下：

The DNAtools interface to ConvertTrace

DNAtools 可以处理“*ConvertTrace*”并且不需要任何使用者干预。将被输入的 trace 文件格式可以被 DNAtools 自动检测并递呈给“*ConvertTrace*”。“*ConvertTrace*”接受大多数普通的 trace 文件格式包括 ABI，ABD 和 SCF。



The DNATools interface to LifeTrace

如何获得 Lifetrace:

Lifetrace 并不包含在 DNAtools 主页的附件中。若用户期望使用 lifetrace 的 Win32 版本, 他需要直接与 [Dirk Walters](#) 联系以获得该程序。

如何安装 Lifetrace:

将 lifetrace.exe 文件放到 DNAtools 的主目录下并启动 DNAtools。

如何在 Linux/Unix 机器上运行 Lifetrace: (略)

If you have access to a computer running Linux or Unix, it is possible to perform the actual base calling on the Linux/Unix machine and subsequent transfer the sequences and the corresponding quality data files to a Windows PC for further processing with dnatools. See [Lifetrace_Linux](#) for further details.

LifeTrace 选项:

Lifetrace 提供一些列选项用于定义 basecalling 的方法和执行序列修剪（在粗序列的起始和结尾部分移除低质量区域）。选项描述如下：

固定的参数用于 basecalling 和修剪：

DNAtools 中的 lifetrace 使用界面包含三个固定的方案用于 basecalling 和修改 trace 文件 (sloppy, standard and stringent)，所有的都基于：

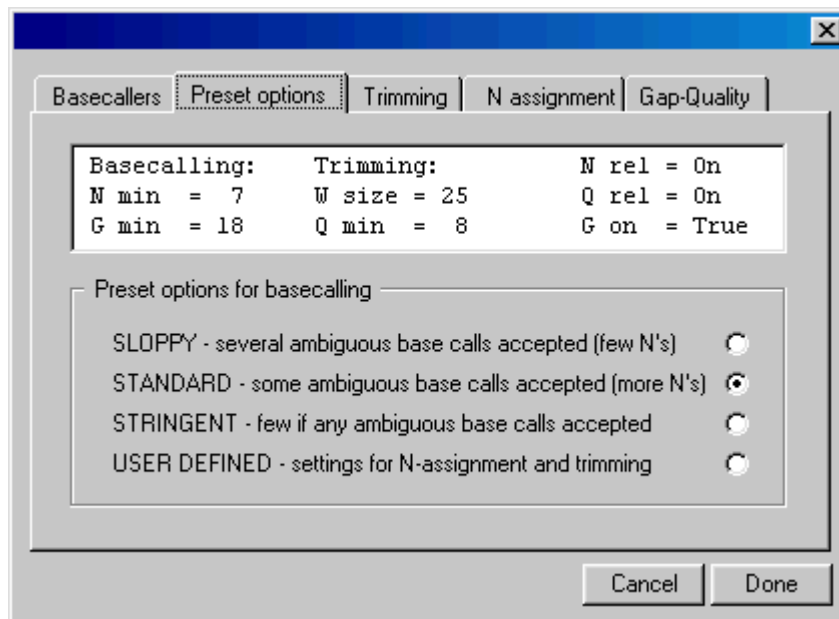
1. 一个最小的 Q 值导致 DNAtools 插入 N 到序列中，而不是一个模棱两可的碱基；
2. 滑行视窗的尺寸（在该视窗中，没有任何碱基的 Q 值低于一个给定的 Q 值）；
3. 在滑行视窗中最小可接受的 Q 值；
4. 最小的 Gap 质量分数，导致 DNAtools 打印 Gap 侧翼的两个碱基。

为了弥补单独的序列 runs 之间的信号强度的差异，已经包含了一个选项用于轻微的修改参数以弥补 trace 文件的信号强度的变化。此项功能激活后，Q 值可以以如下方式进行调整：

$$\text{QualityScore} = \text{QualityScore} + ((\text{MeanScore} - \text{NeutralMean}) / 15)$$

就像上述等式所见的，当所有 trace 文件的质量分数的均值等同于“neutral mean”时，此调整是没有作用的。对于大于“neutral mean”的均数，质量分数终止值增加；而小于“neutral mean”的质量分数终止值则降低。这将允许用户引入一个轻微的调整以弥补单独 trace 文件的全面质量的小变化。

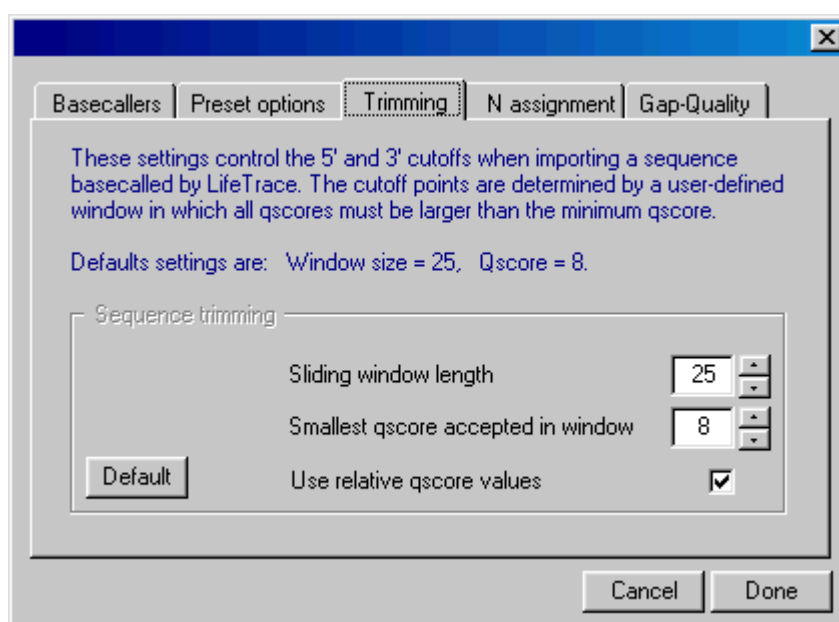
当前选择的用于三个固定设置的这些选项和质量分数、视窗大小和 Gap 质量分数可以被调整（如果被进一步的功能测试建议的话）。当前激活的用于 basecalling 和修剪的参数被显示在“Preset parameters”标号之上的一个文本框中。



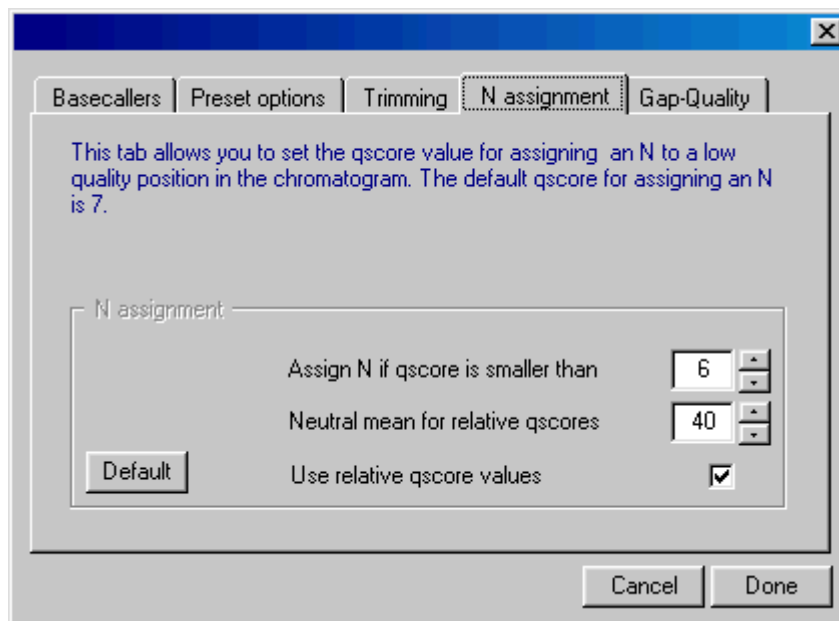
使用者定义参数用于 basecalling 和修改:

为了允许用户优化所有的用于“basecalling 和修改”的参数，接下来的三个表包含用于输入视窗大小的域，用于修改的最小质量分数，用于 N assignment 的质量分数值，用于提示 basecalling 产生的序列中可疑 Gaps 位点的 Gap 质量分数。默认的用于所有使用者定义的设置的参数是和 STANDARD 固定方案中的一样。

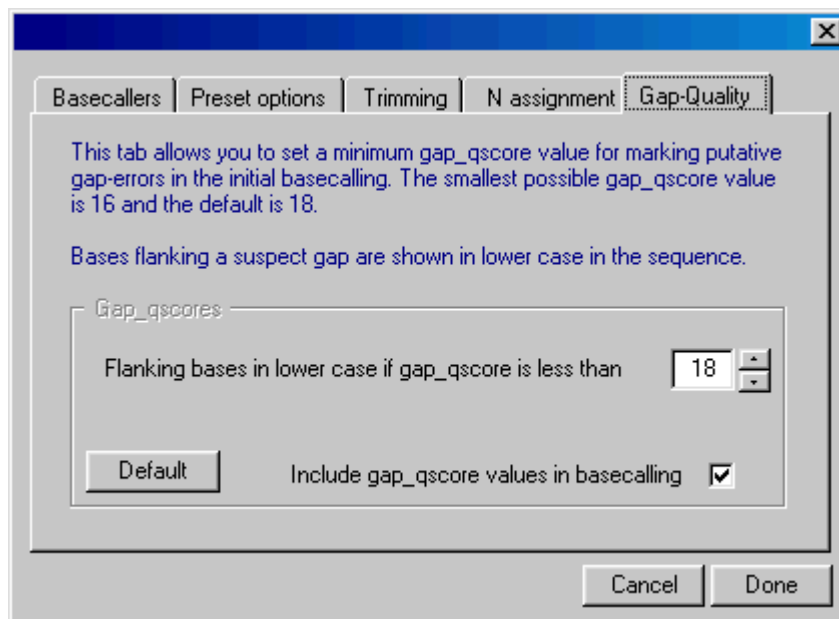
Trimming



N assignment and Neutral mean



Gap 质量分数



6. 运行 DOS 程序:

在视窗下运行 DOS 程序:

这里提及 Clustal W 和双 blast 程序。因为长文件名和目录名是不被 DOS 程序支持的，DNAtools 将 DOS 相关的操作转移到一个分离的目录，DT5_TEMP 定位在 Windows/Winnt

下的一个子目录。当 DNAtools 启动 Clustal 和 Blast 功能时, exe 文件则被拷贝到这个子目录同时 DNAtools 的主目录下的文件和 Blast 数据子目录被删除。

在运行过程中, DOS 程序产生的输入和输出文件被定位到这个目录. 接着 DNAtools 找回这些结果文件用于进一步的加工。用户不用担心 DT5_TEMP 目录, 因为 DNAtools 会自动的移除使用过的文件。

7. 关于文件类型:

DNAtools 使用的文件扩展名:

ALN, PIR, PHY, MSF	Output files from sequence alignment with Clustalw.
DAT, SDF, GCG	Restriction enzyme and user created search data files.
PLP, PSG, PSP	Project path files, used to store the full paths for all files in a project for reloading the complete project or a sub-group of the project.
FOF	File of files. Includes a list including the names of all sequences included in the project.
TXT, RTF, LST, RPT, LOG, TAB	Various ASCII files containing sequence lists, reports, logs etc.
SEQ, DNA, PRO	General extensions for DNA or protein sequence files.
CUT, COD	Codon usage tabels, dnatools and GCG format.
FMS, TMS, MSF, DMS, FAS, FSA	Various types of multi-sequence files.
BLA	Blast search results.
TPL, ESF	Template and complete submission file for transfer of EST sequences to GenBank.
STF, PTF, DTF, MTF, SMF, CGI, TDT	Extensions used in SAGE related functions.
SGD, MCA	Extensions used for files created by EST clustering functions.
OOF, COF, MSG	Primer mail order files.
BMP, WMF	Image files.

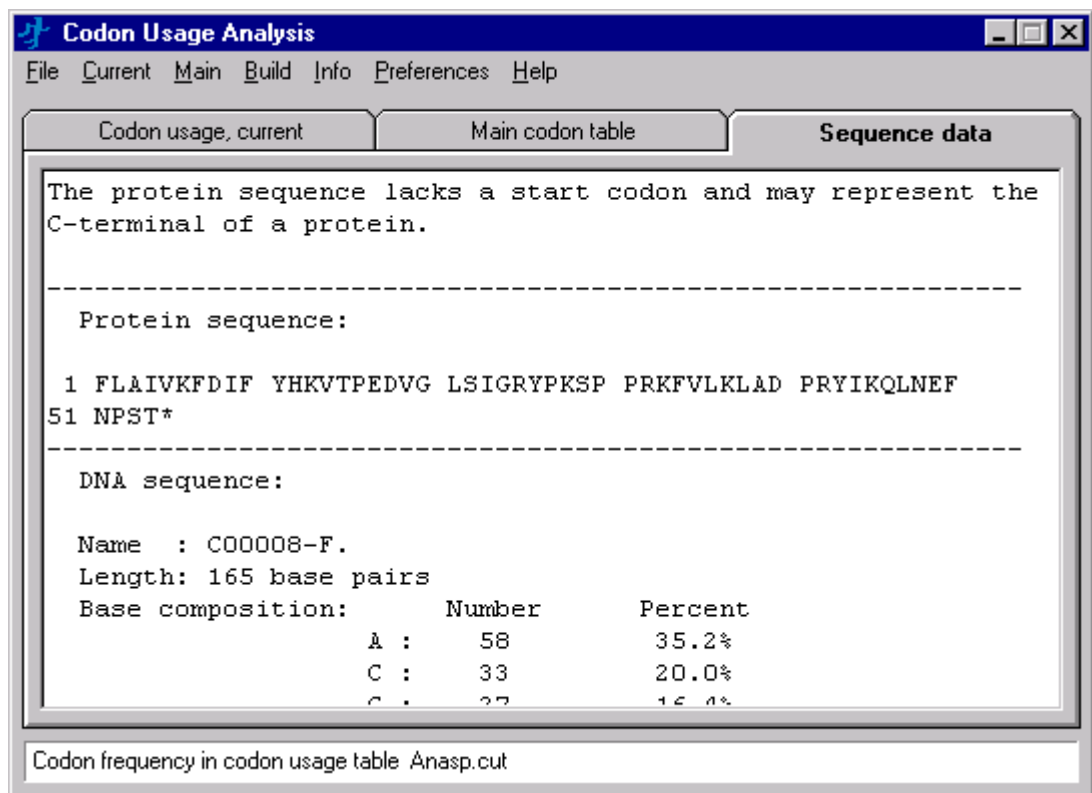
Chapter2: DNAtools-codon usage

1. 密码子使用表:

使用 *Analysis/Codon Usage* 功能用户可以分析一个翻译的 DNA 序列的密码子使用情况。密码子使用数据可以被附加到一个已经存在密码子使用表中，或者用户可以针对一个特殊的有机物通过积累密码子使用数据自己生成数据文件。DNAtools 使用一个不同的格式，但是可以输入和输出 GCG 格式的密码子文件。

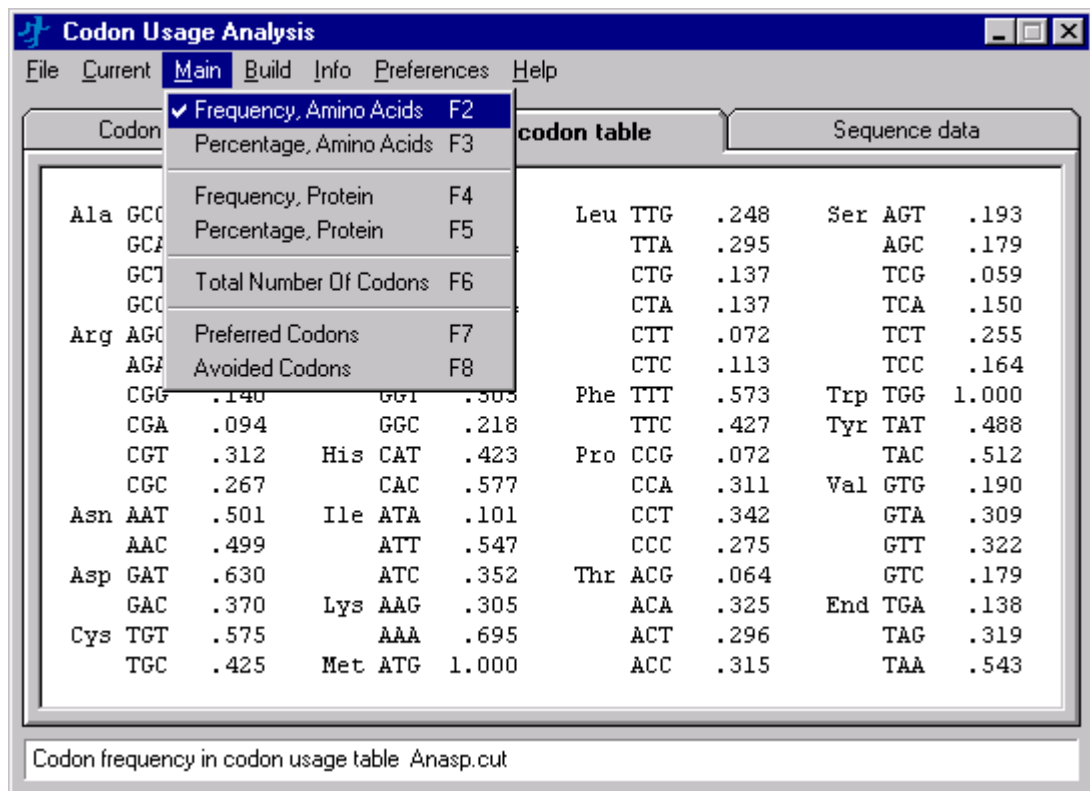
参考 *How to...* 部分寻找更详细的信息以生成一个新的密码子使用表；

该表包含三个域，显示当前序列的密码子使用，主要密码子表，和当前序列的序列数据。



文件菜单包含以下选项：用于打开一个密码子使用表*.cut；输入一个 GCG 格式的密码子使用文件*.cod；以 GCG 格式输出密码子文件*.cod；保存和关闭。

使用主要的密码子使用表来逆向翻译蛋白质序列，以设计 PCR 引物



显示选项：

每个氨基酸的频率；

每个氨基酸的百分率；

整个蛋白质的频率；

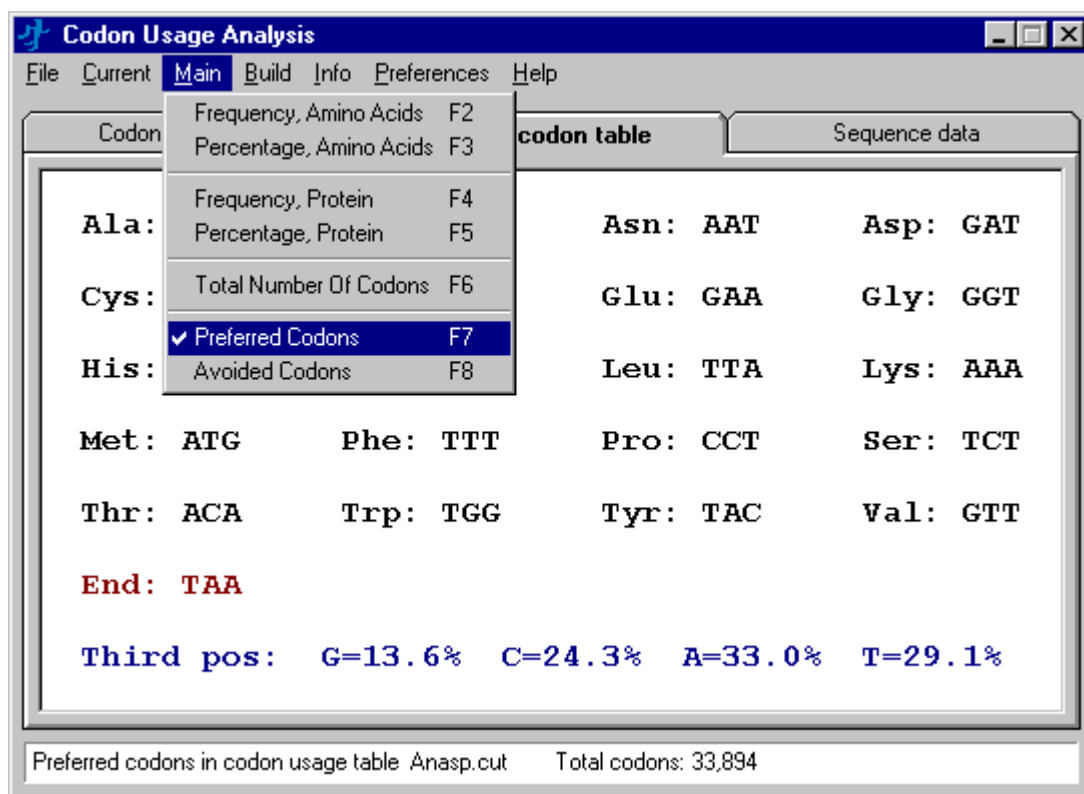
整个蛋白质的百分率；

在蛋白质中密码子总数；

每个氨基酸的优先密码子；

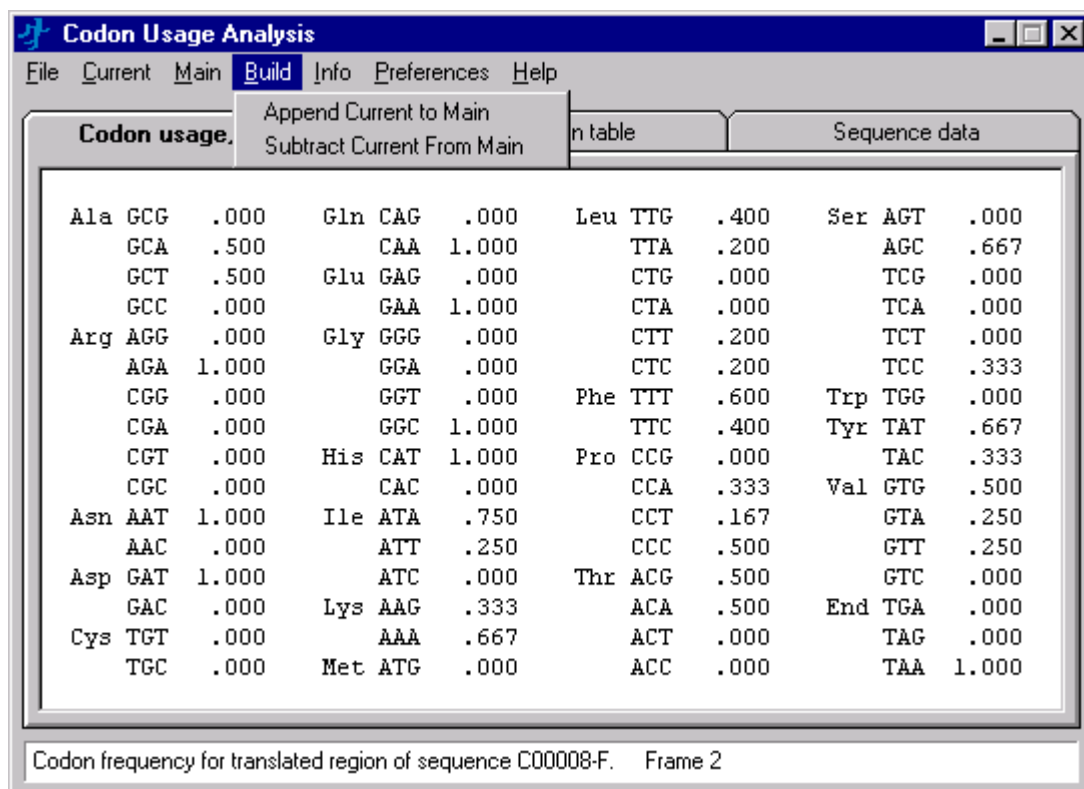
第三位点碱基的百分率；

每个氨基酸避免的密码子；



创建:

允许用户为当前蛋白质序列增加或减少密码子使用数据到主密码子使用表中。在用户可以减少密码子使用数据之前务必显示和翻译想要移除的序列。



信息:

允许用户查看主密码子表的标题和序列的文件名。除非数据集是从包含于主表中文件列表中的序列生成的, 否则不可能从主密码子表中减少数据。

遵循以下步骤生成/修改一个密码子使用表:

从方案中装载用户希望提取密码子使用信息的 DNA 序列;

使用 *Protein/Largest ORF in One Frame /* 翻译序列;

当蛋白质序列展示于编辑器中时, 点击 Format;

DNA 序列和翻译的序列被同时显示在编辑窗口中; 其中翻译的部分被高亮显示;

点击 *Analysis/Codon Usage* 打开密码子表;

点击 *Sequence data* 查看翻译的 DNA 序列;

点击 *Codon usage, current* 查看翻译的 DNA 序列的密码子使用情况;

点击密码子表中的 *File/Open Codon Usage Table* 打开主密码子使用表;

在文件对话框中, 选择已经存在的主表或者打进去一个文件名以生成一个新的主表;

点击 *File/Import GCG Codon File* 可以输入 GCG 格式的主密码子使用表;

主表的内容显示在 *Main codon table* 下;

如果想增加新的密码子使用数据到主表中, 点击 *Build/Append to Main*;

如果想从主表移除新的翻译的序列, 点击 *Build/Subtract*;

关闭密码子使用表。

功能，等等：

Append to Main 附加到主表：当密码子表被展示时，可以为一个蛋白质序列附加密码子信息；

从主表中减少 *Subtract from Main*：只有那些包含于文件列表信息/密码子表文件中的文件可以被从主密码子使用表中移除。

密码子表标题 *Codon table header* -：包含于密码子使用表中的序列的文件名列表被维持在密码子表的标题中；

信息 *Info*：用户可以通过点击主密码子菜单中 info 增加评述到密码子使用文件中。

保存：密码子使用表总是以扩展名为 cut 的格式保存于 DNAtools 中。

文件扩展名：被 DNAtools 创建的密码子使用文件的扩展名为 cut。GCG 格式的密码子使用文件扩展名为 cod。

文件名：当输入时，GCG 格式的文件被转化为 DNAtools 格式且被保存为扩展名为 cut 的文件。

逆向翻译：当用户逆向翻译蛋白质序列时，使用激活的密码子使用文件。

*GCG *.cod files*：输入的 GCG 密码子使用文件缺乏关于序列的起源。因此不可能从这些文件减少密码子数据。

2. 逆向翻译氨基酸序列：

Protein/Back-Translate Protein 命令执行将蛋白质序列翻译成退化的 DNA 序列且同时计算出退化程度。如果退化超过了 10000 次，真实值不被展示。

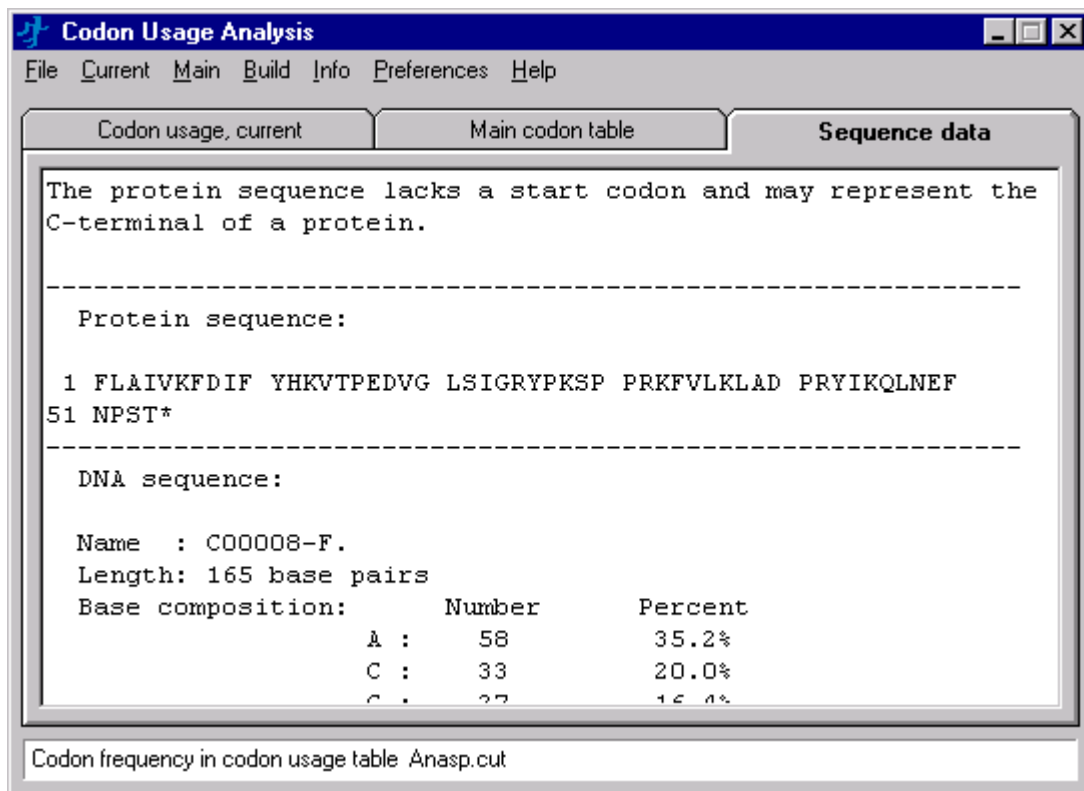
序列验证：

注意不要混淆含 GCG 特征的 DNA 序列和蛋白质序列。如果蛋白肽序列似乎是不正确的话(例如似乎是包含 GCG 特征的 DNA 序列)，程序将提示警告。

自动编辑功能万一回复翻译产生在序列 3' 端有一退化位点的 DNA 序列(如除了 M, ATG 和 W, TGG 之外所有的氨基酸)，包含退化位点的翻译序列将被截断。

密码子使用表:

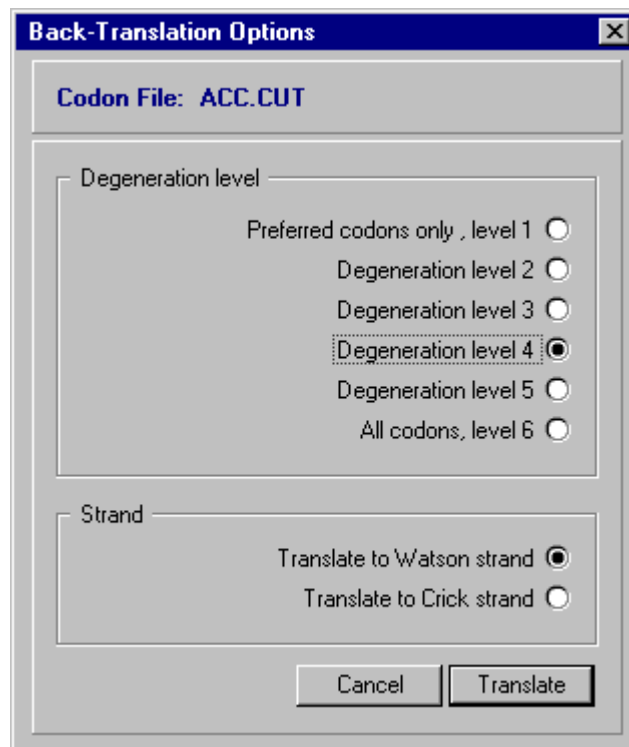
在回复翻译一个蛋白质序列之前，必须从文件菜单中选择密码子格式来装载密码子使用表。



DNAtools 接受两种类型的表，用户生成的或用户修改的*.cut；或者是以 GCG 格式输入的特殊密码子使用表*.cod。后者可以被 DNAtools 简单的转为 DNAtools 的格式，即只需将其另存为*.cut 即可。然而不能保存密码子使用文件为 GCG 格式*.cod。

退化程度:

回复翻译的退化程度可以通过选择退化水平 1—6 进行控制，1 暗示只有首选的密码子才可以用于回复翻译（结果的链是没有退化位点的）。选择 6 意味着可以获得最大退化，同时所有的可能性都包含在链中。2，3，4，5 意思是对每个氨基酸采用最经常使用的密码子。很明显，这将增加链的退化性，只要 2，3，4，5 对于一个给定的氨基酸有不同的密码子。



链：

蛋白质序列既可以被回复翻译成前向链也可被翻译成反向链。对于后者，在计算出退化程度之前回复翻译的序列就被转化为互补的链，同时移除不合规则的 3 端退化位点。

Chapter3: DNAtools-names

1. 关于序列名字：

DNAtools 使用旧的 DOS 文件名。这里解释为什么这样做。

为了全面的利用几个功能（这些功能是为了在一个方案中处理多个文件时而设计的），有必要理解 DNAtools 是如何在这些功能中使用文件名的。这个帮助文档尽力解释为什么在命名序列时相容的行为是这些功能正确的恰当的工作的前提条件。

背景:

在 DNATools 中大多数的与处理多个序列相关的功能是在一个小的执行于 Carlsberg 实验室的 EST 项目中发展而来的。此项目旨在获得关于 Blumeria (霉菌) 的基因组信息和基因的表达情况以更好的理解植物寄生物和其宿主大麦之间的交互作用。

用前向和反向引物, 对所有的来自 cDNA 库并用于分析的克隆测序两次。插入的 5' 序列用于在公开的数据库中进行数据库同源性查找。而 3' polyA 序列则用于生成连接到产生于同样发展阶段的 SAGE。

CDNA 库的插入长度是非常短的(只有很少的是全长的 ORFs), 这反而成为一个优点, 尤其是用于寻找国际性的数据库时。它同时也暗示了在许多场合下, F 和 R 序列是交叠的且可以被这个特殊插入其结合的完全的序列所代替。库的特征允许我们用其结合的序列替换 F 和 R 序列, 这些结合的序列既提高了序列的质量也同时降低了 Blumeria 数据库的序列数据量。

万一某个克隆/插入的 F 和 R 序列并没有交叠, 例如用于连接来自同一克隆的 F 和 R 序列的序列信息是不可获得的, 则使用文件的名字来代替同一插入的 F 和 R 序列的连接。很明显的, 这个需要文件/克隆必须始终按照下面描述的进行命名。

文件名字:

为了跟踪起始于同一插入/插入的 F 和 R 序列, 所有的序列需要依据 DOS 结构命名原则进行命名, 在这种情况下, 一个文件名包含一个八字符名, 一点和三字符扩展名(例如 NAMENAME. EXT)

名字前六个字符表示克隆或插入(e. g. C00018, ABCDEF, 012345).

第七个字符对于 DNAtools 来说不被考虑, 它可以是一个破折号来填满字符数, 或者被用于含同样 F 或 R 引物的多重序列命名, 或者用于引物步行的二级引物命名(e. g. C00018-, ABCDEF-, 012345-, C00018a, ABCDEFa, 012345a)

第八个字符用于确定起始的是 F（前向，沃森，上游链）还是 R（反向，克里克，下游链）（e. g. C00018-F, C00018-R, ABCDEFaF, ABCDEFaR, 012345bF, 012345bR）.

自动的或者是用户整合的 F 和 R 序列接受可以被 DNAtools 识别的特征 M(e. g. C00018-M, ABCDEFaM, 012345bM)。除了 F, R 和 M 之外，在名字的第八个字符是不被 DNAtools 识别的。

三字符的扩展名作为特殊的参数是不被 DNAtools 识别的，但它可以被用于增加额外的信息到文件名中且并不影响多重序列操作。（e. g. C00018-F. seq, C00018-R. old, ABCDEFaF. new）

DNAtools DOS file name: **CCCCCC X S . XXX**

综上：

红色参数对于 DNAtools 有专门的意义（C=克隆名字，S=链），而绿色参数（X=选项）作为特殊的参数是不被 DNAtools 识别的。

长序列名字：

装载一个新的长名字的非 DOS 文件名的序列到 DNAtools 中，DNAtools 会自动的将长文件名传输到 DNAtools 的长文件名变量中。接着长文件名被加工后以产生 DOS 名。最后 DOS 名被传输到 DNAtools 的名字变量中。

对于那些含正确恰当的 DOS 文件名字的新的序列，而这些序列又没有事先被 DNAtools 格式化，一个长文件名会自动的被生成。且这个文件名包含“#”键，后跟 8 个数字。（e. g. C00018-F #47382957）

如果用户希望改变长文件名为一个更加有描述性的序列名，可以通过在常规的标题形式中实现。

长文件名（用于 DNA 和蛋白质和引物序列）包含于所有的印刷物中且和序列文件同时保存。对于引物序列，如果用户想用 DNAtools 进行 E-mail 订购引物时，长文件名是出现于订单形式中唯一的名字。

当行展示选项中定义为行 0 时，在常规的序列标题和主要的编辑形式的信息栏中，长文件名是可见的。看下面，标题行 1—5，用于其他的方法确定展示的序列。

从 DNAtools 不太可能保存含长文件名的序列。

改变文件名：

如果文件名不符合 DNAtools 的准则或缺乏内在的逻辑时，DNAtools 有三个选项用于操纵文件名以帮助用户改变文件名，

如果希望改变已存在的文件名同时保留文件名的起始部分，使用 `change file names`;

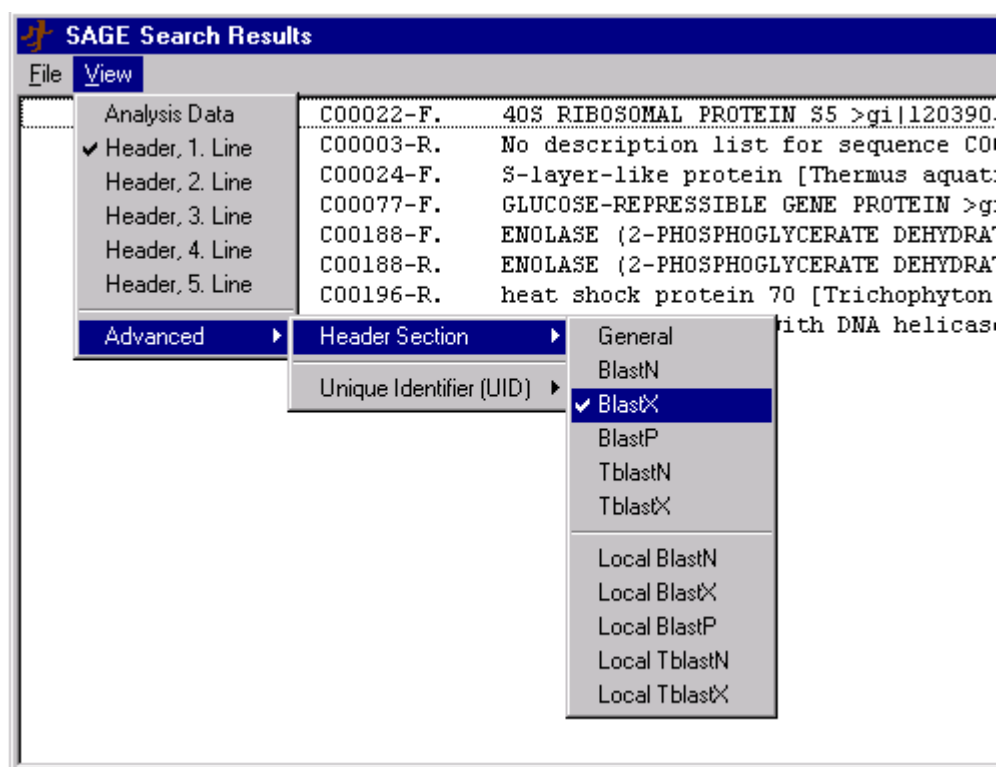
如果希望基于相同的模板生成一个完全新的文件名，使用 `Renaming template`;

如果希望独立的编辑任何一个文件名，使用 `Manual editing`

标题行 1—5：

很明显的，当一个方案包含几千个序列时，不太可能依据 DOS 文件名确定或重新确记一个单个的序列。为了帮助用户记忆，DNAtools 可以使用户功能展示序列中展示序列文本标题的一个选择的行（行 1—5 是指定的引导部分）。使用 View 选项获得菜单结构，显示如下。

选择用户期望使用的引导行，通常行 1 包含最好的数据库匹配。接着去高级选项，选择一个特殊的部分。最后用户可以选择是否希望包含 UID 在展示行中。在大多数情况下，在文件列表中，对于达到“确定”目的，这种做法是没有用的。



如果用户已经执行了（例如，一个 blastx 寻找同时在重新得到寻找结果前也已经审查了自造的选项，将会为每个序列自动的生成一个包含 blastx 描述行的引导部分）。设置 View 菜单选项，显示所有形式以产生序列列表，接着 Blastx 行 1，然后使形式针对所有的序列显示最好的 blastx 数据库匹配。这些使得用户很容易坚定独立的序列，即使该方案包含大量的序列数据

2. 序列名字 手动编辑:

序列名字的手动编辑:

单个独立文件名字的手动编辑可以通过点击在主编辑表中的“显示当前文件名的域”“the field displaying the current file name”来实现。单个独立文件名字的手动编辑应该在重新批命名方案中的所有文件之后才能执行。重新批命名或改变扩展名将会消除事先的任何改变。

文件名核实: 既然手动编辑可能会导致方案中相同文件名的出现，新的文件名将与方案中的所有文件进行比较（初始名字或修改后的名字）。文件名确认还核实是否有不合规定的字符和其他无效的特征。

如果文件名被找到两次或文件名无效，将提示错误信息。为了纠正错误，要么插入初始文件名或继续编辑直致单一文件名被创立。直到单一的新的文件名被接受，否则 DNAtools 的任何功能都不执行。

增加序列入方案：从一个不同的目录下增加新的序列到当前的方案中可能会导致该方案中相同的文件名。但是 DNAtools 会提示警告。

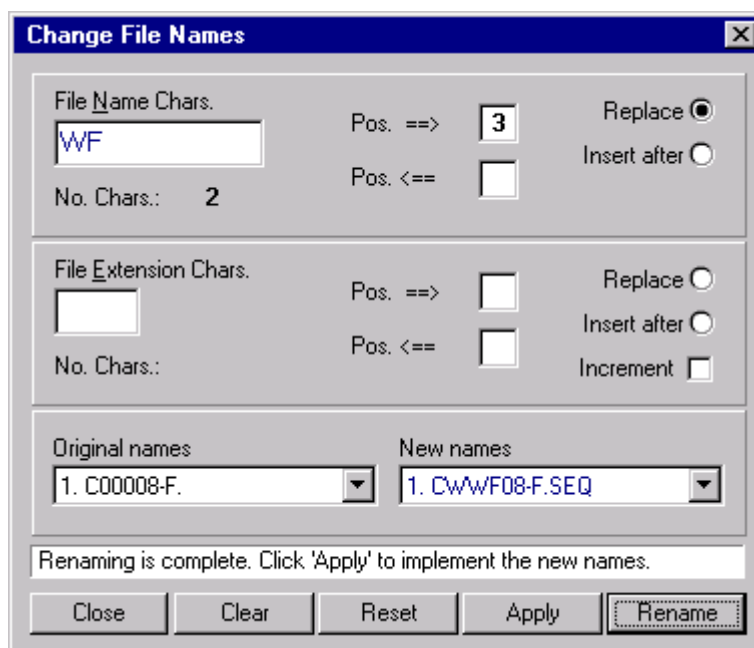
注意：DNAtools 中的用于比较方案中所有序列的几个功能要求依据一般的命令原则构建文件名。

3. 序列名字，批编辑：

此项功能允许用户在不影响名字的其他部分的前提下，对文件名的部分进行复杂的改变。

起初，此项功能将文件名分为 8 字符和 3 字符的扩展名两个部分并且对这两个部分分别对待。运用此项功能可以替换或移除名字中的字符。增加或替换都可从文件名两部分的左边和右边进行。两个 combo 盒展示所有方案中文件的新的名字。

使用这项功能可以移除不需要的字符，但可保旧文件名的其他部分。新的文件名被审查，同时如果有相同的文件名事先存在时重新命名被阻断。



The dialog box is titled "Change File Names" and contains three main sections for configuring file renaming:

- File Name Chars.:** A text input field containing "WF", a "No. Chars.:" label with the value "2", a "Pos. ==>" label with a numeric input "3", a "Pos. <==>" label with an empty input, and radio buttons for "Replace" (selected) and "Insert after".
- File Extension Chars.:** An empty text input field, a "No. Chars.:" label, "Pos. ==>" and "Pos. <==>" labels with empty inputs, and radio buttons for "Replace", "Insert after", and a checked "Increment" checkbox.
- Original names / New names:** Two dropdown menus. The "Original names" dropdown shows "1. C00008-F.". The "New names" dropdown shows "1. CWWF08-F.SEQ".

At the bottom, a status bar reads: "Renaming is complete. Click 'Apply' to implement the new names." Below this are five buttons: "Close", "Clear", "Reset", "Apply", and "Rename" (which is highlighted with a dashed border).

重新命名：依据选项和文本设置，点击命令按钮激活方案中所有文件的重命名功能。如果重命名操作导致相同的文件名，操作被中止并且初始文件名被保留。

清除 N-清除参数以改变文件名；

清除 E-清除参数以改变扩展名；

重新设置-点击命令按钮消除所有的改变。但这并不影响对序列和标题的改变。

关闭-点击命令按钮关闭当前窗口。若不想重新命名，在关闭窗口前重新设置文件名。

文件名字符-文本域可以容纳 8 个字符，这些字符可以被加入到或插入到当前序列名中（通过选项按钮选择）。

扩展名字符-文本域可容纳 3 个字符，这些字符可以被加入到或插入到当前序列名中（通过选项按钮选择）。

位置文件-输入到这些域中的值给出插入的位置或者给出是从文件名和扩展名的左边还是右边进行替换。

增加/插入-文本域/扩展域中的参数可以被增加/插入到文件名/扩展名中（从左从右都可以一按照事先的设定）。插入空格到文件名/扩展名对文件名或扩展名没有影响。若将要被插入的字符数导致文件名或扩展名的长度超过 8 个或 3 个，那些过多的字符会被裁断（从文件名、扩展名的左边或右边）。

替换-文本域扩展域的字符替换文件名或扩展名中同样数量的字符（从左边或右边）。用空格替换字符将会删除文件名或扩展名中的字符。

增加-若文本框只包含数字式字符，将会出现一个核查框。且核查时，导致文本框中的值的增加。

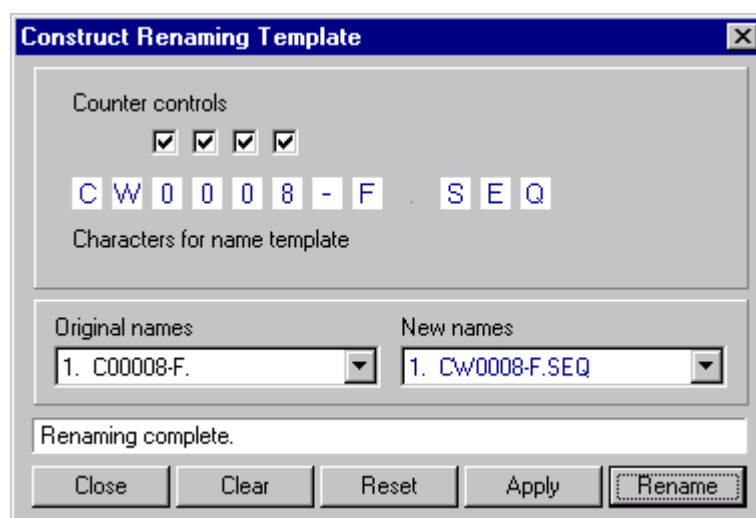
初始名字-展示初始文件名的 Combo 盒。

新的名字-展示新的文件名的 Combo 盒。

注意: DNAtools 中用于执行所有序列比对的一些功能要求依据一般的命名传统构建文件名。

4. 序列名字，用模板进行批创建：

此项功能“Edit/Renaming Template”用于构建一个模板，用此模板重新命名所有当前方案中的序列。模板包含 11 个单字母的文本域（整合在 8 字符的文件名和 3 字符的扩展名中）。若一个文本域包含一个数字，盒顶端将出现一个核查盒。



重新命名一点击该命令按钮激活重新命名所有方案序列的功能。若重新命名模板缺少计算器或若构建的计算器不能提供每个序列一个单独的名字时，程序将提示警告。

还原为初始设置一点击该命令按钮消除任何对序列名的改变。但是对序列和标题的改变无影响。

关闭一点击该命令按钮关闭窗口。若要取消重新命名，首先点击“Reset To Orig”，然后关闭。

计算对照一若输入一个数字式字符到文本域中，在此盒上方出现一个核查盒。两个或更多个邻近的核查盒包含一个计算器（起始于邻近数字给定的值）。计算器的数字和长度是受“文件名的第一和最后部分的可获得长度”限制。单个计算器不能包含所有终止分离分子。计算将起始于邻近盒所指定的值。若没有计算器被激活或者激活的计算器太短以至于不能容纳所有的序列，程序将出现错误。

名字模板的字符一当窗口打开时，所有文件名的共同参数被展示在 11 参数框中(依据他们在文件名中的位置来排列—从文件名的左边界开始计算)。如果共同的字符是一数字，在相关的字符盒的上方将显示一个核查框。没有使用过的参数框是空的。

箭头键，航游一使用左右箭头键定位指针在指定的参数盒中，输入一个字母或一个数字。上下箭头键控制核查框的设置。

初始名字一当装载序列入方案时的一个包含文件名应用的 Combo 盒。

新的名字一包含编辑后的文件名的 Combo 盒。

批重新命名的例子：

蓝色字符是输入的

 X X X X
S E Q 5 0 1 P P . A B 5

1 S E Q 5 0 2 P P . A B 6
2 S E Q 5 0 3 P P . A B 7
3 S E Q 5 0 4 P P . A B 8
4 S E Q 5 0 5 P P . A B 9
5 S E Q 5 0 6 P P . A B 0
6 S E Q 5 0 7 P P . A B 1
7 S E Q 5 0 8 P P . A B 2
8 S E Q 5 0 9 P P . A B 3
9 S E Q 5 1 0 P P . A B 4
10 S E Q 5 1 1 P P . A B 5
11 S E Q 5 1 2 P P . A B 6

注意: DNAtools 中用于执行所有序列比较的几个功能要求依据一般的命名传统构建文件名。

5. 文件和目录工具:

此项功能允许用户打印、保存文件和子目录列表,同时可以修改指定目录下的文件标题。

当用户从 auto-sequencers 受到序列时,后面那项功能是很方便的,只是依据他们在 microtiter 板上的位置被标记。若用户希望移动 trace 文件或序列到另外一个文件夹中,有必要修改文件标题以避免覆盖已经存在的文件。

Example

Plate 01

A01, A02, A03, A04...

B01, B02, B03, B04...

Plate 02

A01, A02, A03, A04...

B01, B02, B03, B04...

With the batch function you can easily add the plate identifier
in-front of all file titles:

Plate 01

01A01, 01A02, 01A03, 01A04...

01B01, 01B02, 01B03, 01B04...

Plate 02

02A01, 02A02, 02A03, 02A04...

02B01, 02B02, 02B03, 02B04...

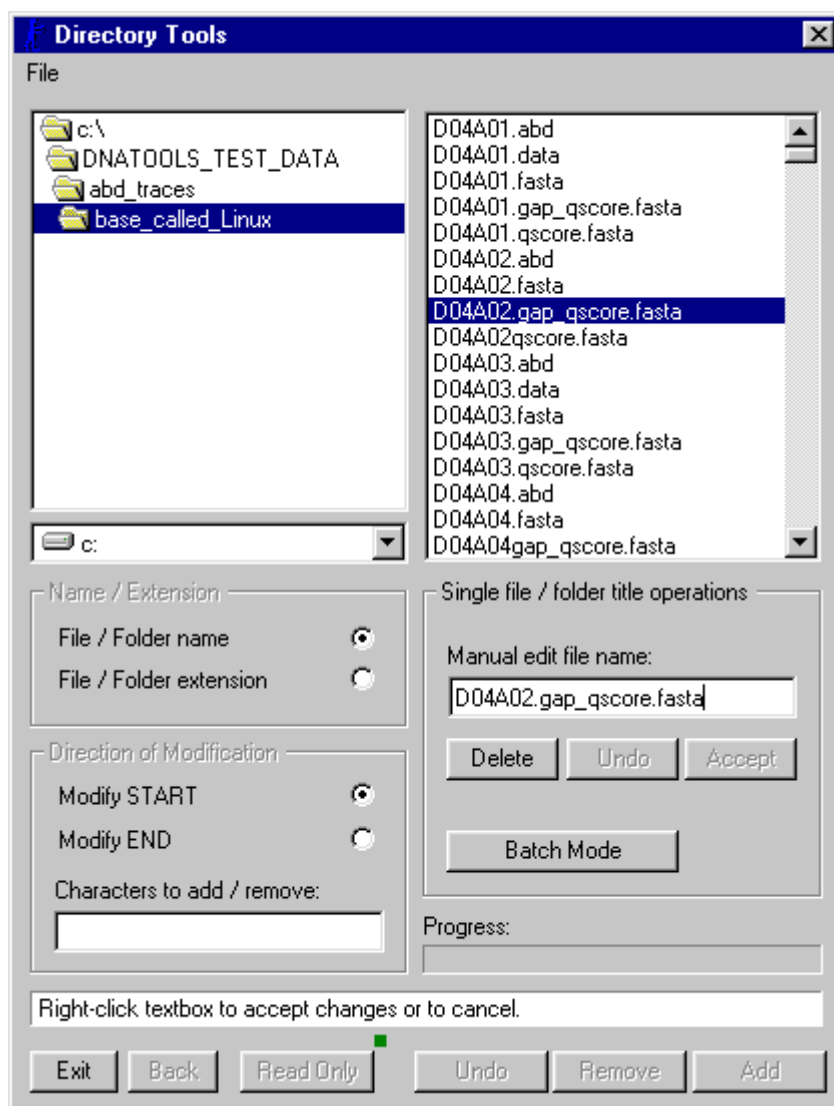
批编辑文件夹中所有文件的标题:

为了增加参数到所有的文件标题中 (= Name.Extension), 首先决定是否希望改变文件名或文件扩展名。然后选择是从文件标题的选定部分的左边还是右边进行改变。最后, 输入希望增加或移除的字符并点击增加或移除按钮。为了从文件名或扩展名中移除参数, 所有文件夹中文件必须包含用户希望移除的参数 (文件标题的右边)。撤销选项允许用户撤销最后十次批改变。

单个文件名的手动编辑:

左键点击选择单个文件或文件中的文件夹标题或文件夹列表, (在文件夹上或在文件列表上)。然后右键点击高亮显示的文件/文件夹标题以拷贝它到编辑域中, 用于手动的编辑或删除。做好改变后点击接受按钮。点击撤销按钮恢复到初始的文件/文件夹标题。为了永久的移除一个文件, 点击 delete。对于此项操作无撤销按钮。

存档/只读按钮设定或移除只读属性。颜色标记提示当前状态, 红色标记提示只读, 而绿色则为存档。



打印和保存文件列表：

允许用户打印和保存列表（名字，大小和日期）—所有文件或和一个指定目录下的亚目录。

警告：

使用此功能时要小心。改变程序的名字或系统文件的名字可能会导致 PC 故障。

6. 序列列表显示选项：

点击编辑器右下角的 L 按钮或者按下 CTRL+L。

列表包含旧的和新的名字，日期和文件最后保存的时间，序列起始和长度和一个序列特异的 5 数字审查总和。

接下来的密码是用于提示序列的起始：WS，沃森链；CS，克里克链；WI，反向沃森链；CI，反向克里克链。

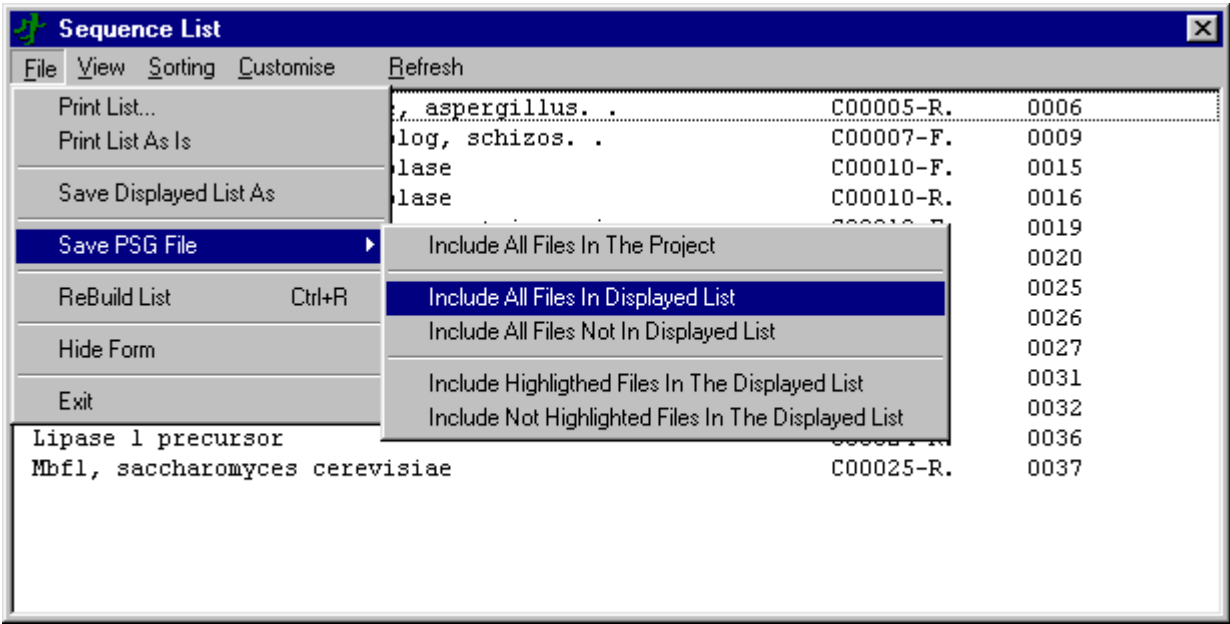
双击列表中的一个项目重新寻回编辑窗口中的选定序列。

文件菜单：

打印列表一点击该项打开打印表。选择完整的文件列表以打印序列数据或标题总结（如果用户希望打印序列标题的 1—5 行）。

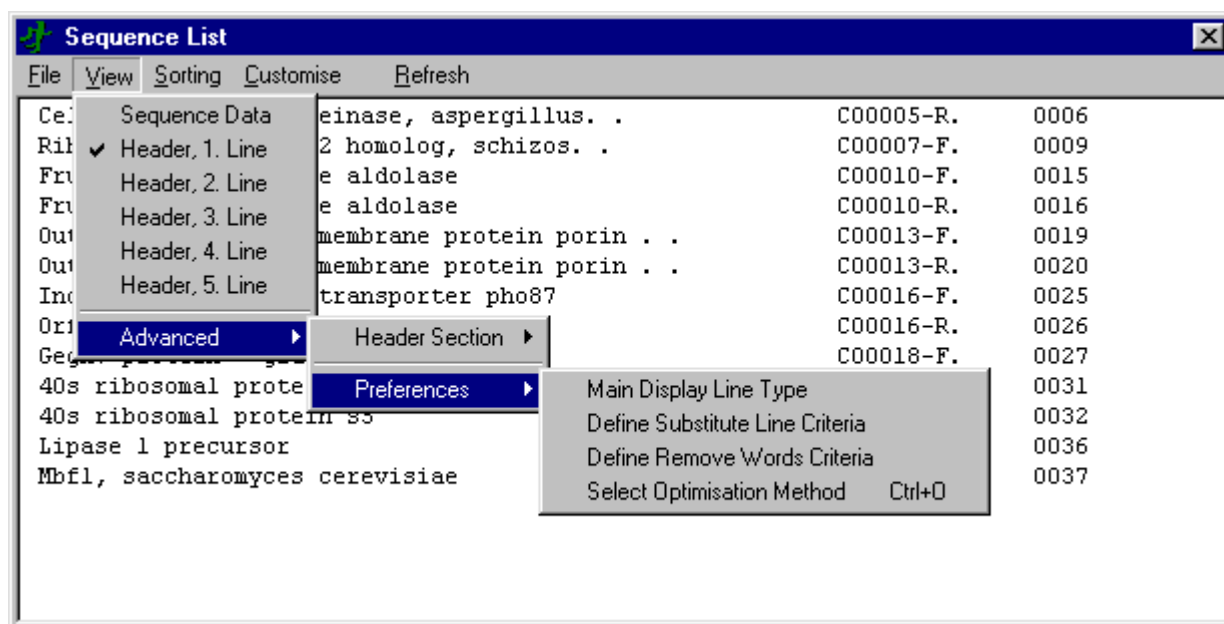
保存为一完整的序列列表可以被保存为一个正常的文本文件。这些文件还可以被输入到其他程序中。

保存 PSG 文件—允许用户保存序列文件的不同的亚组，依据不同的选项设置选定。



查看菜单：

允许用户选定哪种类型的信息用于生成序列列表。如果方案包含超过 700 个序列，用于展示序列列表的列表框容量将过超。 为了避免这些，序列列表的每一行被修短以容纳当前方案中的所有序列。



序列数据—通过状态行查看当前方案中的文件。

标题行 1—5—如果序列标题中的信息是和其他当前方案中所有文件中同一行的同一类型信息一起被系统的排列时，此选项是很有用的。使用 *General* 部分以包含使用者信息同时为同一类型的信息使用同一行。

For example:

- Line 1: Sequence name and origin
- Line 2: Highest homology in Blast search, DNA
- Line 3: Highest homology in Blast search, Protein
- Line 4: User comments

高级选项:

如果序列标题是 DNAtools 自动创建的并且包含 blast 数据库搜索结果、序列和或 Medline 记录，可以使用高级选项来列出序列文件（依据标题特定的日期）。用户可以选择依据一个特定部分（Blastn, Blast x, Blastp, Tblastn or Tblastx）的第一行来列出文件，含或不含 UID(unique record identifier)。

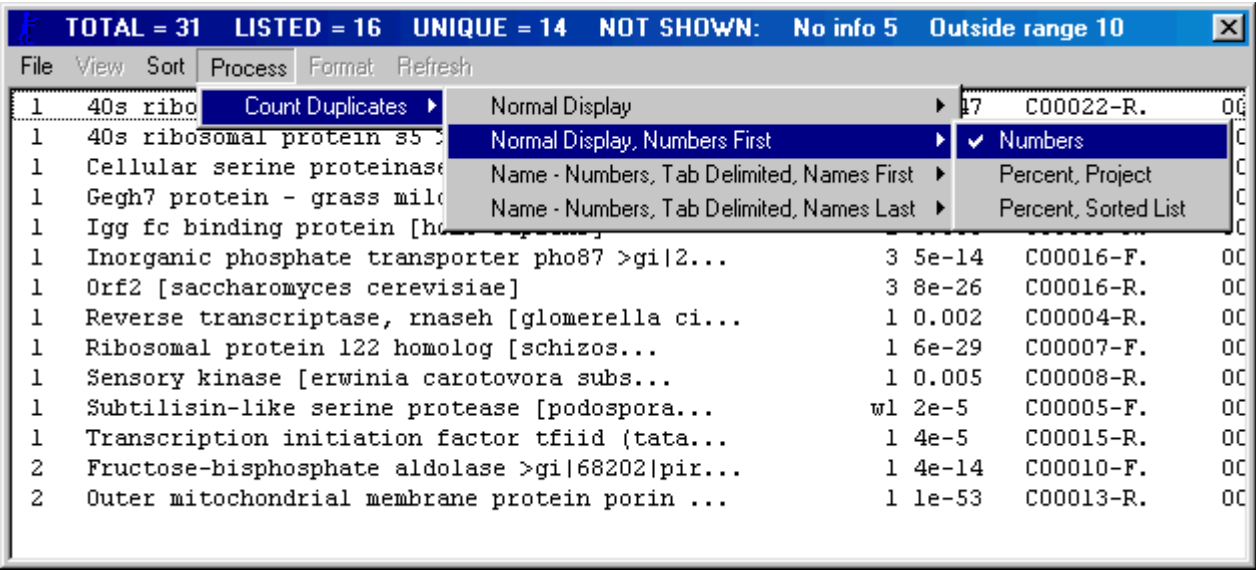
警告：如果用户没有按照自建格式来使用关于标题的高级列表选项，结果将不可预测。

分类菜单：

方案次序—以他们被装载入方案中的序列方式显示序列。

依字母顺序—显示当前方案中的以字母分类的序列。当序列列表是以标题行展示时，此选项尤其有用。这还可以激活加工选项。

加工菜单：



只有当分类序列列表被展示时，此菜单选项才可以用。此项功能基于 blast 搜索结果执行“clustering”，隐藏相同序列，只给出一个给定序列在方案中被找到的次数。这将给用户一个快速的对当前方案中的序列多余性的总结。注意此项功能只考虑名字并不执行任何形式的序列比较。

7. 查看列表参数：

此表允许用户优化文件列表的显示。此处的设置同时可用于“Create FastA Definition Lines, Create EST Submission File, Build SAGE Reliable Mapping Files”和大多数文件列表。在选择好指定的选项后，使用主表上的 L 按钮显示优化结果。三个例子如下：

用于格式化文件列表的选项允许用户优化单独行的内容，借助于这些行以展示序列。注意：在序列标题中无 blast 数据，将没有东西格式化并且行展示只限于序列名字，长度等。

注意：文件列表包含文件菜单下的一些选项，可以允许用户处理一些或所有列出的序列（或通过保存一个方案亚组 (*.psg)，或打开一个新的包含选定序列亚组的 DNAtools 实例，或保存选定的序列到多序列文件中 (MS, in fasta format)）。序列列表中选项有：

1. All sequences included in the project
2. All sequences included in the list
3. All sequences NOT included in the list
4. All highlighted sequences in the list
5. All sequences in the list NOT highlighted.

使用这些选项可以隔离几乎任何文件/序列亚组以用于进一步的分析。

加工序列列表：

与序列列表有关的一个新的特征是在文件列表上的加工选项。使用该功能，用户可以通过移除相同序列和仅仅一次列出序列和其发生在行中的次数来优化一个分类序列列表（当序列是依据他们在方案中的数字而列出的，此功能不能使用）。

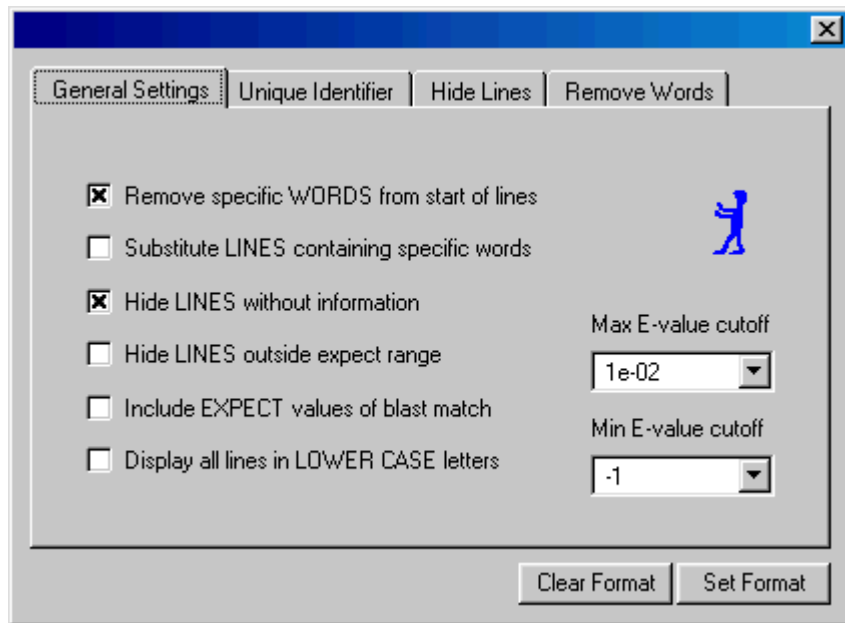
因为选项是非常广泛的，用户得进行一次小规模尝试以获得用户期望获得的展示。

列表查看选项：

在此表中可以选择不同的列表选项。注意打开 “*Hide lines with e-value of blast match worse than cutoff*”也可激活“*Hide lines without blast information in header section*”。

一般设置：

从下拉列表中选择期望终止值被用于程序的每个部分以区别重要的和不重要的 blast 比对。选定的值被用于所有的 blast 程序选择。若用户希望使用不同的终止值用于 blastn 比对，用户必须改变优先选择表中的值。



展示行类型：

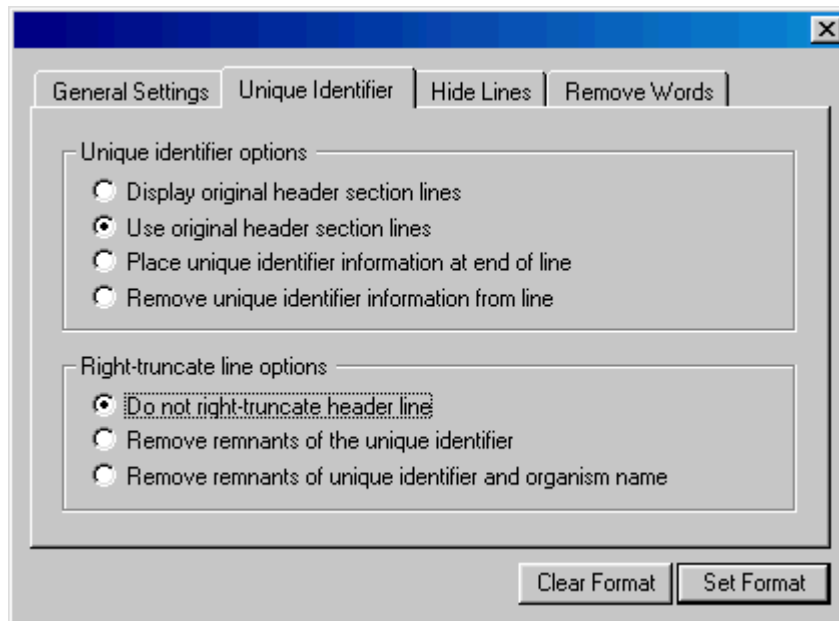
此表的上部被用于以下四种主行展示类型之一：

1. No modification at all of lines, except that lines are truncated a bit at the end to fit the form.
2. Unmodified lines with accession numbers in front. Line substitution and word removal active
3. Moves the unique identifier the end of the line. Line substitution and word removal active
4. Completely removes unique identifier. Line substitution and word removal active

当列表是按照字母分类的，选项 1—3 是相关的。注意：若选项 1 是被选择的。

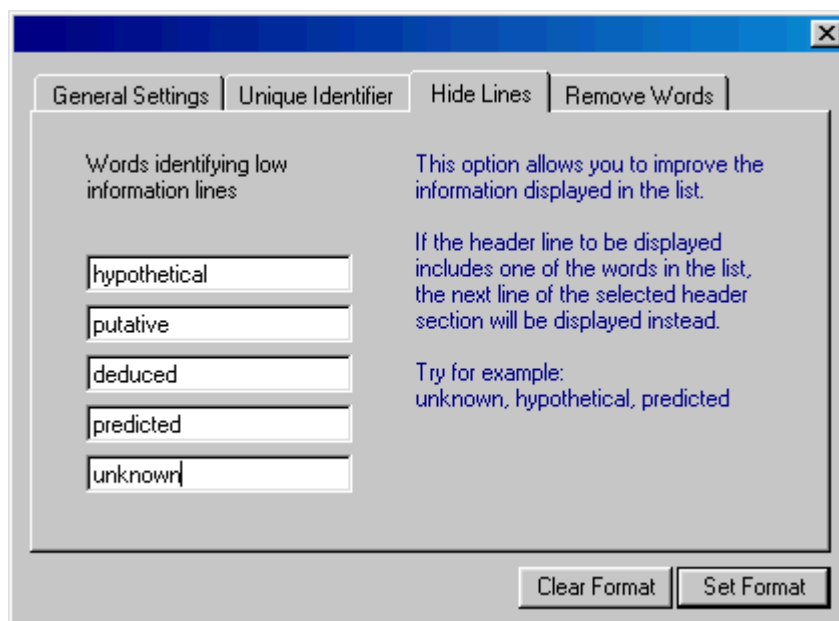
右截短行：

因为从 blast 程序返回的描述行经常被截短，行有时包含单独标志符的和或有机体名字的残余部分。通过选择选项 2(only UID remnants)或选项 3(UID remnants and name of the organism)从行的右端移除这些残余部分。



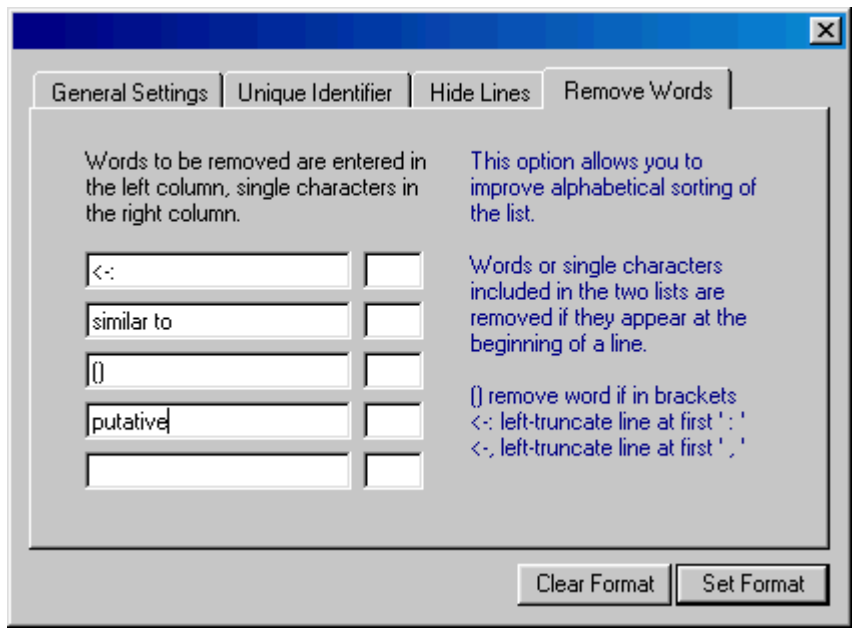
隐藏行标准:

在单词列表中最多可以输入 5 个单词。列表中的单词被用于鉴定含低信息内容的行。若选定的标题部分的第一行包含一个或更多个单词，程序将检测下一行以确定一个或更多个单词的存在。直到 blastx 对比指定的终止值或者含一个可接受的 blastx 比对的行更差时此行为才终止。真正展示的描述行的数量在行中显示。



移除单词列表:

在单词列表中最多可以输入 5 个单词，在字符列表中最多可以输入 5 个单个的字符。若该选项被激活，用这些单词或字符的一个起始的行被左截短，因此单词或字符被移除。对于提高基因名字的分类是有用的。对于那些单词或字符被移除的行，程序用 W 或 C 提示。输入 “()” 强迫移除括号中的领导单词。“<-:” 暗示：若在行的前 15 字符找到 “:”，行是被左截短的。选项对于 “，” 同样适用。



Examples:

未修改的文件列表:

列表的每一行是完整的第一描述行（来自安字母分类的 blastx 搜索）:

```
*No information in header line
1 C00006-R.
*No information in header line
1 C00009-F.
*No information in header line
1 C00018-R.
*No information in header line
1 C00019-F.
*No information in header line
1 C00042-F.
```


*No information in header line			
1			C00050-F.
*No information in header line			
1			C00051-R.
*No information in header line			
1			C00053-R.
*No information in header line			
1			C00056-F.
*No information in header line			
1			C00057-R.
*No information in header line			
1			C00058-R.
*No information in header line			
1			C00062-R.
*No information in header line			
1			C00068-R.
*No information in header line			
1			C00070-R.
gi 1166378 (L76169) reverse transcriptase, RNaseH [Glomerella			
ci. . 42 0.002			C00004-R.
gi 1561655 (M76953) envelope glycoprotein [Human			
immunodeficienc. . 31 3.9			C00051-F.
gi 1654344 (U64105) guanine nucleotide exchange factor			
p115-RhoG. . 38 0.041			C00053-F.
gi 171554 (M81879) galactose permease [Saccharomyces			
cerevisiae] 31 5.3			C00059-F.
gi 173189 (M21089) growth regulation protein [Saccharomyces			
cere. . 50 7e-06			C00052-F.
gi 2253267 (AF005668) properdin [Homo			
sapiens] 31 5.0			C00032-F.
gi 2262187 (U56099) FacB [Aspergillus			
niger] 30 7.6			C00062-F.
gi 2262187 (U56099) FacB [Aspergillus			
niger] 36 0.12			C00031-F.
gi 2408009 (Z99161) ubiquitin [Schizosaccharomyces			
pombe] 50 7e-13			C00061-R.
gi 2454622 (AF018033) reverse transcriptase [Magnaporthe			
grisea] 34 0.34			C00075-R.
gi 2612882 (AF015825) NodB-like protein [Bacillus subtilis]			
>gi . . 38 0.039			C00033-R.
gi 2622767 (AE000923) cell division control protein Cdc48			
[Metha. . 32 1.3			C00048-R.
gi 2707187 (U94183) unknown [Glomerella			
cingulata] 89 8e-19			C00025-R.

gi|2905804 (AF047689) subtilisin-like serine protease
[Podospora. . 49 2e-05 C00005-F.

gi|4039014 (AF037338) cleft lip and palate transmembrane
protein. . 98 3e-20 C00037-R.

gi|4079831 (AF083464) DNA polymerase zeta catalytic subunit
[Mus. . 34 0.61 C00050-R.

gi|516040 (U12335) cAMP-dependent protein kinase catalytic
subun. . 107 3e-23 C00059-R.

gi|669031 (U20864) similar to glycerol facilitator protein and
o. . 31 3.0 C00057-F.

gi|732944 (X82499) protein kinase [Saccharomyces
cerevisiae] 59 2e-08 C00047-R.

gi|870734 (L43065) suppresses the respiratory deficiency of a
ye. . 33 0.76 C00060-F.

gi|993026 (X90369) S-layer-like protein [Thermus aquaticus
therm. . 30 6.4 C00024-F.

gnl|PID|d1003374 (D13716) 'multicopy suppressor of stt4
mutation. . 34 0.59 C00014-R.

gnl|PID|d1012779 (D83776) The KIAA0191 gene is expressed
ubiquit. . 30 6.6 C00070-F.

gnl|PID|d1013761 (D86349) ribosomal protein L22 homolog
[Schizos. . 38 0.038 C00007-R.

gnl|PID|d1013761 (D86349) ribosomal protein L22 homolog
[Schizos. . 104 6e-29 C00007-F.

gnl|PID|d1020288 (D84239) IgG Fc binding protein [Homo
sapiens] 40 0.008 C00009-R.

gnl|PID|d1026325 (AB012663) polyprotein [Azuki bean mosaic
virus] 31 5.4 C00008-F.

gnl|PID|d1032513 (AB016006) ribosomal protein S31 homolog
[Schiz. . 107 6e-23 C00046-F.

gnl|PID|e1187364 (Y12504) BiP protein [Aspergillus
awamorii] 123 4e-28 C00028-F.

gnl|PID|e1264048 (AJ224971) DNA polymerase [Feline herpesvirus
1] 29 9.1 C00015-F.

gnl|PID|e1284411 (AL022244) hypothetical protein
[Schizosaccharo. . 107 7e-34 C00016-F.

gnl|PID|e1284411 (AL022244) hypothetical protein
[Schizosaccharo. . 124 3e-28 C00016-R.

gnl|PID|e1314599 (AL031187) putative transposable element
[Arabi. . 35 0.26 C00023-R.

gnl|PID|e1316202 (Y14067) ferrichrome-iron receptor
[Salmonella . . 31 4.1 C00065-R.

gnl|PID|e1319411 (AL031530) putative nadh-cytochrome b5
reductas. . 106 6e-23 C00040-F.

gnl|PID|e1319411 (AL031530) putative nadh-cytochrome b5
 reductas. . 165 2e-40 C00040-R.
 gnl|PID|e1343600 (Z81458) C03E10.3 [Caenorhabditis
 elegans] 31 3.1 C00046-R.
 gnl|PID|e1349397 (Z66524) Homology with Squid retinal-binding
 pr. . 74 4e-13 C00037-F.
 gnl|PID|e267541 (X98931) heat shock protein 70 [Emericella
 nidul. . 31 4.2 C00036-R.
 gnl|PID|e311463 (Z93940) unknown [Bacillus subtilis]
 >gi|2145382. . 31 3.9 C00075-F.
 gnl|PID|e319086 (Y13338) cellular serine proteinase
 [Aspergillus. . 132 9e-31 C00005-R.
 gnl|PID|e339894 (Z98956) Nad5 protein [Dicranum
 scoparium] 30 5.5 C00006-F.
 gnl|PID|e351516 (Y13670) sensory kinase [Erwinia carotovora
 subs. . 41 0.005 C00008-R.
 pir||I52657 gene SEZ-6 (seizure-related 6) protein precursor -
 m. . 32 1.3 C00066-F.
 pir||JC4749 gEgh7 protein - grass mildew >gi|727157 (L40638)
 gEg. . 107 3e-23 C00018-F.
 pir||R3BY31 ribosomal protein S25.e.A precursor, cytosolic -
 yea. . 111 2e-24 C00068-F.
 pir||S23313 hypothetical protein 1 - Arabidopsis thaliana
 retrot. . 35 0.27 C00023-F.
 pir||S23988 ubiquitin / ribosomal protein CEP52 - fruit fly
 (Dro. . 175 2e-43 C00061-F.
 pir||S43743 probable dual specificity phosphatase (EC 3.1.3.-)
 -. . 37 0.039 C00065-F.
 pir||S65236 probable membrane protein YPL217c - yeast
 (Saccharom. . 143 5e-36 C00058-F.
 pir||S70765 nodulin-45 precursor - narrow-leaved blue
 lupine 34 0.57 C00048-F.
 sp|013695|YEN1_SCHPO HYPOTHETICAL 52.9 KD SERINE-RICH PROTEIN
 C1. . 30 8.0 C00041-F.
 sp|P03756|VE22_LAMBD EA22 GENE PROTEIN >gi|76109|pir||ZEBP2L
 Ea2. . 114 5e-35 C00072-F.
 sp|P03756|VE22_LAMBD EA22 GENE PROTEIN >gi|76109|pir||ZEBP2L
 Ea2. . 238 2e-62 C00072-R.
 sp|P06100|NOT2_YEAST GENERAL NEGATIVE REGULATOR OF
 TRANSCRIPTION. . 67 7e-11 C00045-F.
 sp|P06215|CHIT_PHAVU ENDOCHITINASE PRECURSOR >gi|169331
 (M13968). . 53 8e-07 C00033-F.
 sp|P06262|NU4C_TOBAC NADH-PLASTOQUINONE OXIDOREDUCTASE CHAIN
 4, . . 33 1.0 C00066-R.

sp|P07144|PORI_NEUCR OUTER MITOCHONDRIAL MEMBRANE PROTEIN
 PORIN . . 166 7e-41 C00013-F.
 sp|P07144|PORI_NEUCR OUTER MITOCHONDRIAL MEMBRANE PROTEIN
 PORIN . . 208 1e-53 C00013-R.
 sp|P07902|GAL7_HUMAN GALACTOSE-1-PHOSPHATE URIDYLTRANSFERASE
 >. . 34 0.44 C00031-R.
 sp|P14540|ALF_YEAST FRUCTOSE-BISPHOSPHATE ALDOLASE
 >gi|68202|pir. . 78 4e-14 C00010-F.
 sp|P14540|ALF_YEAST FRUCTOSE-BISPHOSPHATE ALDOLASE
 >gi|68202|pir. . 122 4e-37 C00010-R.
 sp|P15690|NUAM_BOVIN NADH-UBIQUINONE OXIDOREDUCTASE 75 KD
 SUBUNI. . 30 5.0 C00012-R.
 sp|P17671|E75A_DROME ECDYSONE-INDUCIBLE PROTEIN E75-A
 >gi|103149. . 30 6.6 C00019-R.
 sp|P20261|LIP1_CANRU LIPASE 1 PRECURSOR >gi|422081|pir||S23448
 t. . 83 1e-15 C00024-R.
 sp|P22151|GRG1_NEUCR GLUCOSE-REPRESSIBLE GENE PROTEIN >gi|3014
 (. . 71 4e-12 C00077-F.
 sp|P26783|RS5_YEAST 40S RIBOSOMAL PROTEIN S5 (RP14) (YS8)
 >gi|10. . 130 1e-47 C00022-R.
 sp|P27941|AMC2_ORYSA ALPHA-AMYLASE ISOZYME C2 PRECURSOR
 (1,4-ALP. . 35 0.26 C00045-R.
 sp|P32323|AGA1_YEAST A-AGGLUTININ ATTACHMENT SUBUNIT PRECURSOR
 >. . 33 0.75 C00047-F.
 sp|P34237|YKR9_YEAST HYPOTHETICAL 77.5 KD PROTEIN IN PRP1-STE3
 I. . 45 3e-04 C00035-R.
 sp|P34675|YO25_CAEEL HYPOTHETICAL 202.6 KD PROTEIN ZK688.5 IN
 CH. . 32 2.2 C00032-R.
 sp|P47140|YJ70_YEAST HYPOTHETICAL 37.5 KD PROTEIN IN YUH1-URA8
 I. . 31 4.2 C00035-F.
 sp|P49687|N145_YEAST NUCLEOPORIN NUP145 (NUCLEAR PORE PROTEIN
 NU. . 29 8.7 C00014-F.
 sp|P78695|GR78_NEUCR 78 KD GLUCOSE REGULATED PROTEIN HOMOLOG
 PRE. . 103 6e-22 C00028-R.
 sp|P87207|MNT3_CANAL PROBABLE MANNOSYLTRANSFERASE MNT3
 >gi|21903. . 30 8.7 C00042-R.
 sp|Q00495|KFMS_RAT MACROPHAGE COLONY STIMULATING FACTOR I
 RECEPT. . 31 5.0 C00012-F.
 sp|Q03112|EVI1_HUMAN ECOTROPIC VIRUS INTEGRATION 1 SITE
 PROTEIN . . 31 4.1 C00060-R.
 sp|Q12731|TF2D_EMENI TRANSCRIPTION INITIATION FACTOR TFIID
 (TATA. . 47 4e-05 C00015-R.
 sp|Q24186|RS5_DROME 40S RIBOSOMAL PROTEIN S5 >gi|1203905
 (U48394. . 92 2e-18 C00022-F.

优化的文件列表：

1. 在每一行的起始部分，独特的标志符信息已经被移除；
2. 在“替换”列表中包含单词的第一描述行已被接下来的行所代替（优于终止限制值的 blastx 匹配），同时不再包含任何禁止单词。展示的描述行的行数在列表中提及；
3. 如果单词在移除列表中，行的第一个单词则已经被移除。若一个单词或字符已被移除，将有 W 或 C 提示这种变化；
4. 列表是按照字母分类的。

```
*No information in header line 1
C00006-R.      0005
*No information in header line 1
C00009-F.      0010
*No information in header line 1
C00018-R.      0025
*No information in header line 1
C00019-F.      0026
*No information in header line 1
C00042-F.      0051
*No information in header line 1
C00050-F.      0061
*No information in header line 1
C00051-R.      0064
*No information in header line 1
C00053-R.      0067
*No information in header line 1
C00056-F.      0068
*No information in header line 1
C00057-R.      0070
*No information in header line 1
C00058-R.      0072
*No information in header line 1
C00062-R.      0080
*No information in header line 1
C00068-R.      0086
*No information in header line 1
C00070-R.      0088
202.6 kd protein zk688.5 in ch. .
C00032-R.      0040
```

w1 2.2

37.5 kd protein in yuhl-ura8 i. .	w1	4.2
C00035-F. 0043		
40s ribosomal protein s5	1	1e-47
C00022-R. 0029		
40s ribosomal protein s5	1	2e-18
C00022-F. 0028		
52.9 kd serine-rich protein cl. .	w1	8.0
C00041-F. 0050		
77.5 kd protein in prpl-ste3 i. .	w1	3e-04
C00035-R. 0044		
78 kd glucose regulated protein homolog pre. .	1	6e-22
C00028-R. 0036		
A-agglutinin attachment subunit precursor	1	0.75
C00047-F. 0057		
Alpha-amylase isozyme c2 precursor	1	0.26
C00045-R. 0054		
Bip protein, aspergillus awamorii	1	4e-28
C00028-F. 0035		
C03e10.3, caenorhabditis elegans	1	3.1
C00046-R. 0056		
Camp-dependent protein kinase catalytic subun. .	1	3e-23
C00059-R. 0074		
Cell division control protein cdc48, metha. .	1	1.3
C00048-R. 0060		
Cellular serine proteinase, aspergillus. .	1	9e-31
C00005-R. 0003		
Cleft lip and palate transmembrane protein. .	1	3e-20
C00037-R. 0047		
Dna polymerase zeta catalytic subunit, mus. .	1	0.61
C00050-R. 0062		
Dna polymerase, feline herpesvirus 1	1	9.1
C00015-F. 0020		
Dual specificity phosphatase	w1	0.039
C00065-F. 0081		
Ea22 gene protein	1	2e-62
C00072-R. 0090		
Ea22 gene protein	1	5e-35
C00072-F. 0089		
Ecdysone-inducible protein e75-a	1	6.6
C00019-R. 0027		
Ecotropic virus integration 1 site protein . .	1	4.1
C00060-R. 0076		
Endochitinase precursor	1	8e-07
C00033-F. 0041		

Envelope glycoprotein, human immunodeficienc. .	1	3.9
C00051-F. 0063		
Facb, aspergillus niger	1	7.6
C00062-F. 0079		
Facb, aspergillus niger	1	0.12
C00031-F. 0037		
Ferrichrome-iron receptor, salmonella . .	1	4.1
C00065-R. 0082		
Fructose-bisphosphate aldolase	1	4e-14
C00010-F. 0012		
Fructose-bisphosphate aldolase	1	4e-37
C00010-R. 0013		
Galactose permease, saccharomyces cerevisiae	1	5.3
C00059-F. 0073		
Galactose-1-phosphate uridylyltransferase	1	0.44
C00031-R. 0038		
Gegh7 protein - grass mildew	1	3e-23
C00018-F. 0024		
Gene sez-6	1	1.3
C00066-F. 0083		
General negative regulator of transcription. .	1	7e-11
C00045-F. 0053		
Glucose-repressible gene protein	1	4e-12
C00077-F. 0093		
Glycerol facilitator protein and o. .	w1	3.0
C00057-F. 0069		
Growth regulation protein, saccharomyces cere. .	1	7e-06
C00052-F. 0065		
Guanine nucleotide exchange factor p115-rhog. .	1	0.041
C00053-F. 0066		
Heat shock protein 70, emericella nidul. .	1	4.2
C00036-R. 0045		
Homology with squid retinal-binding pr. .	1	4e-13
C00037-F. 0046		
Igg fc binding protein, homo sapiens	1	0.008
C00009-R. 0011		
Inorganic phosphate transporter pho87	3	5e-14
C00016-F. 0022		
Lipase 1 precursor	1	1e-15
C00024-R. 0033		
Macrophage colony stimulating factor i recept. .	1	5.0
C00012-F. 0014		
Mannosyltransferase mnt3	w1	8.7
C00042-R. 0052		

Mbf1, <i>saccharomyces cerevisiae</i>	3	1e-07
C00025-R. 0034		
Membrane protein ypl217c - yeast	w1	5e-36
C00058-F. 0071		
Multicopy suppressor of stt4 mutation. .	c1	0.59
C00014-R. 0019		
Nad5 protein, <i>dicranum scoparium</i>	1	5.5
C00006-F. 0004		
Nadh-cytochrome b5 reductas. .	w1	2e-40
C00040-R. 0049		
Nadh-cytochrome b5 reductas. .	w1	6e-23
C00040-F. 0048		
Nadh-plastoquinone oxidoreductase chain 4, . .	1	1.0
C00066-R. 0084		
Nadh-ubiquinone oxidoreductase 75 kd subuni. .	1	5.0
C00012-R. 0015		
Nodb-like protein, <i>bacillus subtilis</i>	1	0.039
C00033-R. 0042		
Nodulin-45 precursor - narrow-leaved blue lupine	1	0.57
C00048-F. 0059		
Nucleoporin nup145	1	8.7
C00014-F. 0018		
Orf2, <i>saccharomyces cerevisiae</i>	3	8e-26
C00016-R. 0023		
Outer mitochondrial membrane protein porin . .	1	1e-53
C00013-R. 0017		
Outer mitochondrial membrane protein porin . .	1	7e-41
C00013-F. 0016		
Polyprotein, azuki bean mosaic virus	1	5.4
C00008-F. 0008		
Properdin, <i>homo sapiens</i>	1	5.0
C00032-F. 0039		
Protein 1 - <i>arabidopsis thaliana</i> retrov. .	w1	0.27
C00023-F. 0030		
Protein kinase, <i>saccharomyces cerevisiae</i>	1	2e-08
C00047-R. 0058		
Reverse transcriptase, <i>magnaporthe grisea</i>	1	0.34
C00075-R. 0092		
Reverse transcriptase, <i>rnaseh</i> , <i>glomerella ci.</i> .	1	0.002
C00004-R. 0001		
Ribosomal protein 122 homolog, <i>schizos.</i> .	1	0.038
C00007-R. 0007		
Ribosomal protein 122 homolog, <i>schizos.</i> .	1	6e-29
C00007-F. 0006		

Ribosomal protein s25.e.a precursor, cytosolic - yea. .	1	2e-24
C00068-F. 0085		
Ribosomal protein s31 homolog, schiz. .	1	6e-23
C00046-F. 0055		
Sensory kinase, erwinia carotovora subs. .	1	0.005
C00008-R. 0009		
S-layer-like protein, thermus aquaticus therm. .	1	6.4
C00024-F. 0032		
Subtilisin-like serine protease, podospora. .	1	2e-05
C00005-F. 0002		
Suppresses the respiratory deficiency of a ye. .	1	0.76
C00060-F. 0075		
The kiaa0191 gene is expressed ubiquit. .	1	6.6
C00070-F. 0087		
Transcription initiation factor tfiid	1	4e-05
C00015-R. 0021		
Transposable element, arabi. .	w1	0.26
C00023-R. 0031		
Ubiquitin / ribosomal protein cep52 - fruit fly	1	2e-43
C00061-F. 0077		
Ubiquitin, schizosaccharomyces pombe	1	7e-13
C00061-R. 0078		
Unknown, bacillus subtilis	1	3.9
C00075-F. 0091		

优化的文件列表:

此优化与上面显示的一样, 除了: 低于终止值的 blastx 匹配的序列已经从列表中被移除了。

40s ribosomal protein s5	1	1e-47
C00022-R. 0029		
40s ribosomal protein s5	1	2e-18
C00022-F. 0028		
78 kd glucose regulated protein homolog pre. .	1	6e-22
C00028-R. 0036		
Bip protein, aspergillus awamorii	1	4e-28
C00028-F. 0035		
Camp-dependent protein kinase catalytic subun. .	1	3e-23
C00059-R. 0074		
Cellular serine proteinase, aspergillus. .	1	9e-31
C00005-R. 0003		
Cleft lip and palate transmembrane protein. .	1	3e-20
C00037-R. 0047		

Ea22 gene protein	1	2e-62
C00072-R. 0090		
Ea22 gene protein	1	5e-35
C00072-F. 0089		
Endochitinase precursor	1	8e-07
C00033-F. 0041		
Fructose-bisphosphate aldolase	1	4e-14
C00010-F. 0012		
Fructose-bisphosphate aldolase	1	4e-37
C00010-R. 0013		
Gegh7 protein - grass mildew	1	3e-23
C00018-F. 0024		
General negative regulator of transcription. .	1	7e-11
C00045-F. 0053		
Glucose-repressible gene protein	1	4e-12
C00077-F. 0093		
Growth regulation protein, saccharomyces cere. .	1	7e-06
C00052-F. 0065		
Homology with squid retinal-binding pr. .	1	4e-13
C00037-F. 0046		
Inorganic phosphate transporter pho87	3	5e-14
C00016-F. 0022		
Lipase 1 precursor	1	1e-15
C00024-R. 0033		
Mbf1, saccharomyces cerevisiae	3	1e-07
C00025-R. 0034		
Membrane protein ypl217c - yeast	w1	5e-36
C00058-F. 0071		
Nadh-cytochrome b5 reductas. .	w1	2e-40
C00040-R. 0049		
Nadh-cytochrome b5 reductas. .	w1	6e-23
C00040-F. 0048		
Orf2, saccharomyces cerevisiae	3	8e-26
C00016-R. 0023		
Outer mitochondrial membrane protein porin . .	1	1e-53
C00013-R. 0017		
Outer mitochondrial membrane protein porin . .	1	7e-41
C00013-F. 0016		
Protein kinase, saccharomyces cerevisiae	1	2e-08
C00047-R. 0058		
Ribosomal protein l22 homolog, schizos. .	1	6e-29
C00007-F. 0006		
Ribosomal protein s25.e.a precursor, cytosolic - yea. .	1	2e-24
C00068-F. 0085		

Ribosomal protein s31 homolog, schiz. .	1	6e-23
C00046-F. 0055		
Subtilisin-like serine protease, podospora. .	1	2e-05
C00005-F. 0002		
Transcription initiation factor tfiid	1	4e-05
C00015-R. 0021		
Ubiquitin / ribosomal protein cep52 - fruit fly	1	2e-43
C00061-F. 0077		
Ubiquitin, schizosaccharomyces pombe	1	7e-13
C00061-R. 0078		

Chapter4: DNAtools-compare sequence

1. 本地 blast 搜索:

这个功能允许用户可以在本地数据库执行搜索（如果用户希望搜索大于一个的当地数据库，参考本地多数据库搜索）。至于 clustal 多序列队列功能，这个功能使用外部的 DOS 程序来执行真正的搜索，但是操作界面却是 DNAtools 提供的。

在用户进行 blast 搜索之前,Formatdb 和 Blastall 程序必须被正确的安装,以允许 DNAtools 可以与 DOS 程序对话。用户还需要依据自己的序列或抽取自公共数据库的序列来生成一个或更多的本地数据库。

如何建立 DNAtools 用于本地的 blast 搜索:

从 DNAtools 的 ftp 站点(<ftp://ftp.crc.dk/pub/dnertools/blastz.exe>) 或从 NCBI(<ftp://ncbi.nlm.nih.gov/blast/executables/blastz.exe>) 下载压缩的 blastz.exe 文件(2,878 Kb, version 2.0.9, May 07, 1999)。

拷贝压缩的文件到 DNAtools 的目录下。确信在 DNAtools 的 Data 目录下不存在子目录。如果有，移除或删除它。然后运行 blastz.exe。这将会在 Data 目录下生成一个新的子目录和 5 个 exe 文件: blastall.exe, formatdb.exe, blastpgp.exe, fastacmd.exe 和 seedtop.exe。

当用户下次启动 DNAtools 时，如果 Data 目录, formatdb.exe 和 blastall.exe 存在的话，它会自动的启动 blast 功能并且在“搜索”菜单中(blastall)和“使用/

多序列功能”中 *Utilities/Multi-Sequence Functions* (formatdb)展示这些功能。
用户随时准备生成自己的数据库同时用 5 种 blast 程序的任意程序搜索他们 (blastn, blastx, blastp, tblastn, tblastx)。

在启动之后，所有 blast 文件将被从 DNAtools 目录中移除。

手动/方案搜索：

点击“手动”命令按钮以隐藏方案选项，并且展示一个文本域以用于手动进入或经过一个寻求行。一个 DNA 行必须至少是 8 个碱基，而蛋白质行则至少是 4 个氨基酸残基。

手动搜索：

Blast Local Search.

AAAAAATATATTTCAAGTTATATTCATCCTGGCGCATTACG
TACAACGGAGATGAGTGTTAAATAAAATAAAACAATTTT
ATTTTACGATCCTTGGCAGCAAAGAACATTGTCACCTACCTC
CCGACAGATAGCATGTTGTAGATAGGGCGATTTCATTGTGA
ATATGGAATAGCCAATGATCCACTGCGGAGACTCGATGGGA
TCGACC

Get Seq. Clear 211 residues Project

Blast settings

Descriptions 5 Program BlastN
Alignments 0 Database BLUMERIA
Expect limit 1e-4 Both strands of query ☒
Filter query sequence ☐

Progress 0%

Enter a sequence and click 'Search'.

Close Stop Search

点击“方案”命令按钮隐藏手动选项，同时允许用户选择包含于当前方案中的序列。

自动搜索一个方案：

Blast Local Search.

Range, Unknown project type

All ☐ Search from sequence

Current ☐ Search to sequence

Number of sequences

Destination of result

View result, don't insert in header ☒

Append result to header section ☐

Replace header section ☐

Get Seq. Manual

Blast settings

Descriptions Program

Alignments Database

Expect limit Both strands of query ☒

Filter query sequence ☐

Progress

Select sequence(s) from the project and click 'Search'.

Close Stop Search

当本地 blast 搜索表被打开时，如果当前展示序列的一个区域被选择或加亮的话，则选定的区域被转移到这个表中的手动搜索域中。点击这个表获得 seq 有相同的作用。

在开始搜索之前，DNAtools 检查询问序列的类型、blast 程序和数据库同时提示用户是否参数不相容。如果用户坚持进行搜索的话，程序将马上启动。

输出结果的加工：

为了对包含于方案中的序列进行本地的 blast 搜索，对“结果的命运” (*Destination of result*) 的设置决定了如何处理搜索结果。结果既可以被插入到已经存在的序列标题的前面，或者直接替换当前的标题。

警告：对于后一种情况，存在于本地 blast 标题部分的当前信息将会被覆盖，而且没有提示警告。且如果用户保存了这个方案的话，这些信息将不可逆的丢失。

对于手动搜索，尽管在一个方案中可以进行一个范围内的序列搜索，但是结果只能观看。搜索结果必须被保存在序列标题中。

在视窗下运行 DOS 程序：

这里提及 Clustal W 和双 blast 程序。因为长文件名和目录名是不被 DOS 程序支持的，DNAtools 将 DOS 相关的操作转移到一个分离的目录，DT5_TEMP 定位在 Windows/Winnt 下的一个子目录。当 DNAtools 启动 Clustal 和 Blast 功能时，exe 文件则被拷贝到这个子目录同时 DNAtools 的主目录下的文件和 Blast 数据子目录被删除。

在运行过程中，DOS 程序产生的输入和输出文件被定位到这个目录。接着 DNAtools 找回这些结果文件用于进一步的加工。用户不用担心 DT5_TEMP 目录，因为 DNAtools 会自动的移除使用过的文件。

Example of search output for blastn search on ERYSHIPHE data base:

BLASTN 2.0.9 [May-07-1999]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database
search
programs", Nucleic Acids Res. 25:3389-3402.

Query= C00045-F. Ori: WS Len: 402 Chk: 23262
(402 letters)

Database: C:\WINNT\DT5_TEMP\ERYSHIPHE
4524 sequences; 1,883,240 total letters

```

Score E
Sequences producing significant alignments: (bits) Value

C00045-F Len: 402 Check: 23262 797 0.0
C00240-R Len: 384 Check: 72676 551 e-158
D00496-R Len: 365 Check: 20475 44 4e-005
D00110-F Len: 486 Check: 88899 44 4e-005
C00491-F Len: 292 Check: 2468 44 4e-005

>C00045-F Len: 402 Check: 23262
Length = 402

Score = 797 bits (402), Expect = 0.0
Identities = 402/402 (100%)
Strand = Plus / Plus

Query: 1
aaaatgccaatTTCAATGAGGAGCCTTATTATTTATGTTTTAACAGCAACCCTGGGACCT 60
|||||
|||||
Sbjct: 1
aaaatgccaatTTCAATGAGGAGCCTTATTATTTATGTTTTAACAGCAACCCTGGGACCT 60

Query: 61
tcagcaggttatggcggcacaggaactgttggttatgccctcctgaatgattcagctcta 120
|||||
|||||
Sbjct: 61
tcagcaggttatggcggcacaggaactgttggttatgccctcctgaatgattcagctcta 120

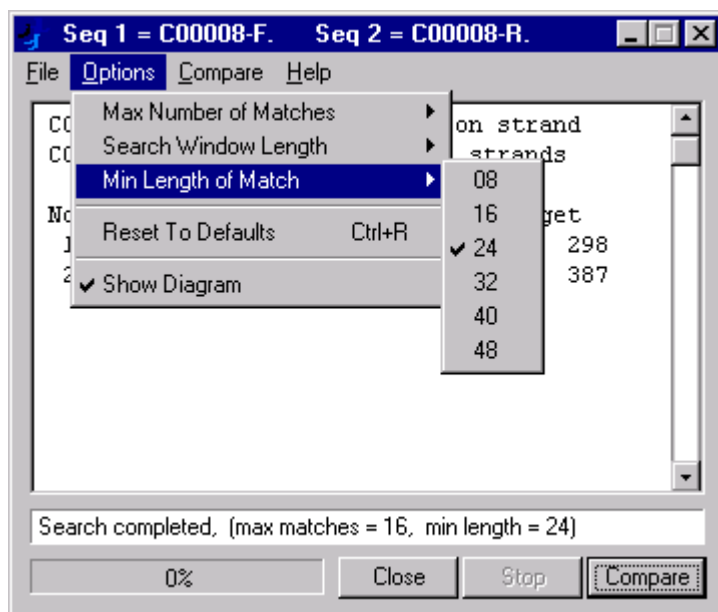
```

2. 比较两个序列 相同的区域:

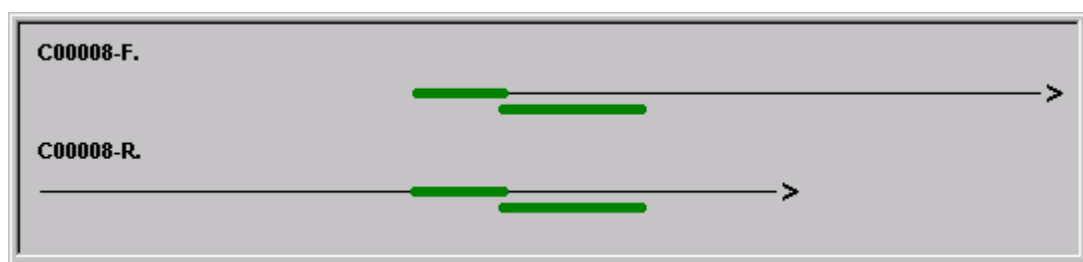
“*Search/Compare Two Sequences*”命令的比较是在选自于文件列表中的两个序列中进行的。

如果没有选择序列，当前序列将和其自身比较。搜索可在沃森，克里克或两种链上进行。通过改变搜索选项可以修改搜索的敏感性和速度。

结果可以显示为分类的匹配列表，并给出相同区域的长度和相同的部分。W 提示匹配是针对沃森链而 C 提示匹配是针对相互不的询问链。所有的相配之物都涉及当前链和询问链的沃森链。



另外，两个序列被展示在图谱中，并显示出相同的部分。



注意：这个功能只检测两个序列之间完全相同的部分。

文件菜单：

选择序列：打开文件列表以选择用于比较的两个序列。按下 CTRL 键允许用户选择两个序列。第一个序列是源序列，而第二个序列则是目标序列。选择额外的序列没有任何作用。只要一个序列被选择了，它会自身进行比较以揭示内在的相同性。（如直接重复或反向重复）。

从列表框中选择一个或两个序列之后，按下回车键完成选择并返回到比较表。

打印：打印两个序列中相同部分的列表。在每个匹配的前面加上 W（沃森链）或 C（克里克链）提示匹配的起始部分与当前序列的关系。打印结果包含一个标题和一个页脚。

选项：

匹配最大数：定义匹配的最大数（5—50）限制搜索参数“搜索相同的最大区域”。

定义一个较小的参数值将会增加搜索速度。

视窗长度：定义搜索视窗长度（2—20）。一个小值将产生更敏感的搜索，当搜索时间延长。

最小匹配长度：定义显示在结果列表中的最小相同区域（3—30）

显示图表：展示源序列和目标序列队列的图表。两个序列被队列排列以达到最长的相同区域匹配。绿色片断表示在源序列和目标序列之间的相同部分。在源序列中的红色片断和目标序列中蓝色片断是互补的。

比较：

沃森链：和询问链的沃森链进行相同性搜索。

克里克链：和询问链的克里克链进行相同性搜索。询问序列的相配之物指的是克里克链。

两条链：和询问链的两条链进行相同性搜索。询问序列的相配之物指的是沃森链（W）或克里克链（C）。

中止：直接行动

立即取消同源性搜索并清除视窗。如果比较花了过分长的时间，使用这个选项。增加搜索视窗的长度和或降低匹配数将减少搜索时间。

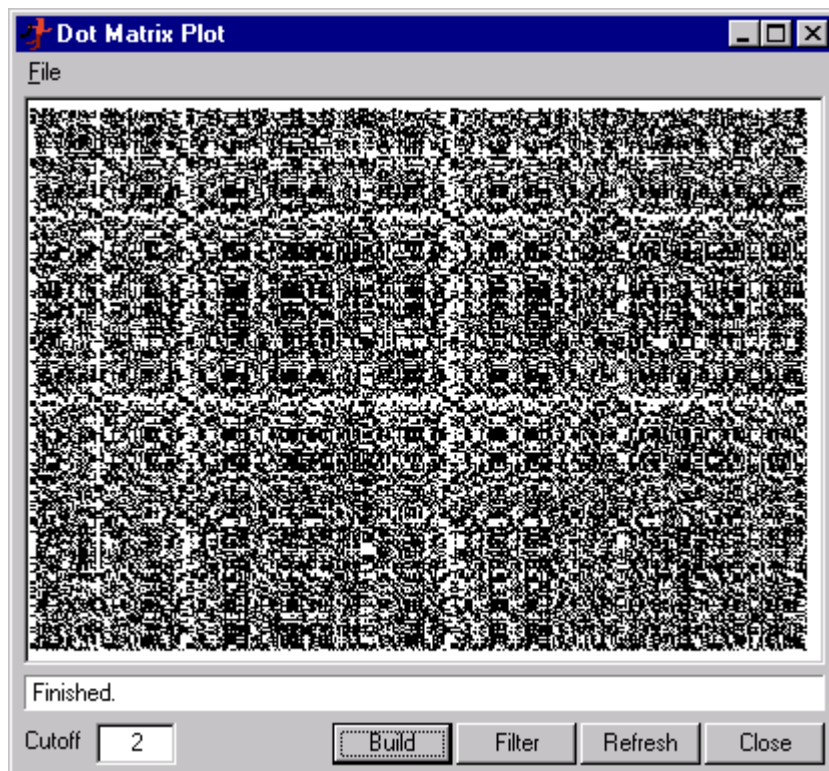
3. 比较两个序列 点阵

这项功能允许用户比较两个序列，或者和序列自身比较，通过使用点阵列的方法来寻找两个序列中相同的区域。分析过程包括两步：第一步，需要一个完整的矩阵（序列 1 的长度乘以序列 2 的长度），这个矩阵包含布尔数学体系信息（真或假），且

这些信息是用于比较序列 1 中的所有碱基和序列 2 中的所有碱基。在下面的点图中，所有的配对以点表示，而无配对的则不表示。

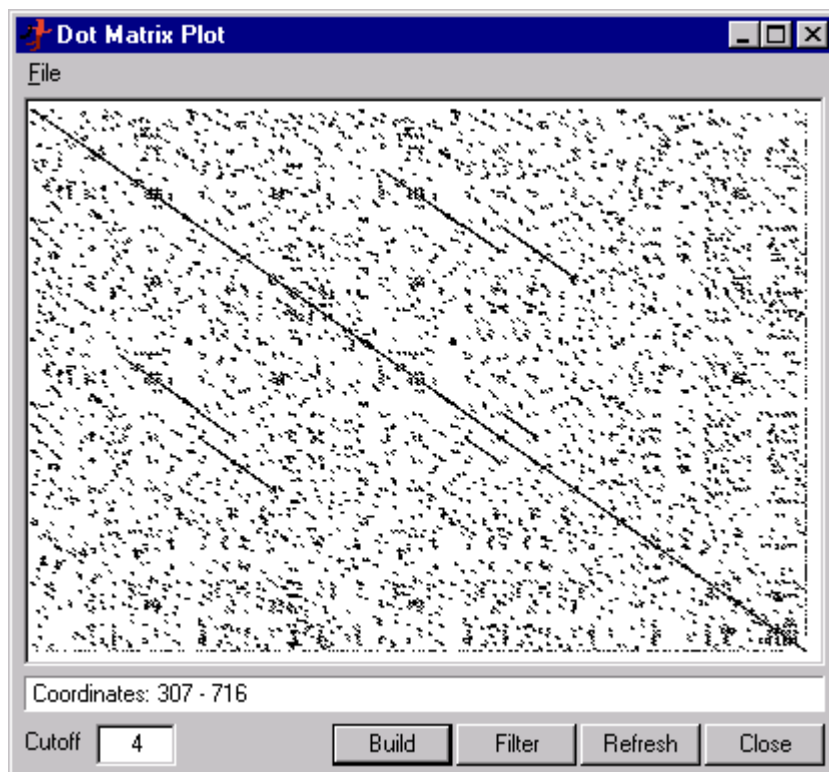
在比较序列之前，当用户点击“文件/选择序列”时，这些序列必须从展示的文件列表中选择。在选择一个或两个序列之后，点击“建造”以生成完整的点阵。对于长序列，可能需要一些时间。当矩阵完成时，点击“过虑”以清晰的展示矩阵。表大小可变，但是没有被刷新的话自己是不会重画的。

Complete Dot Matrix

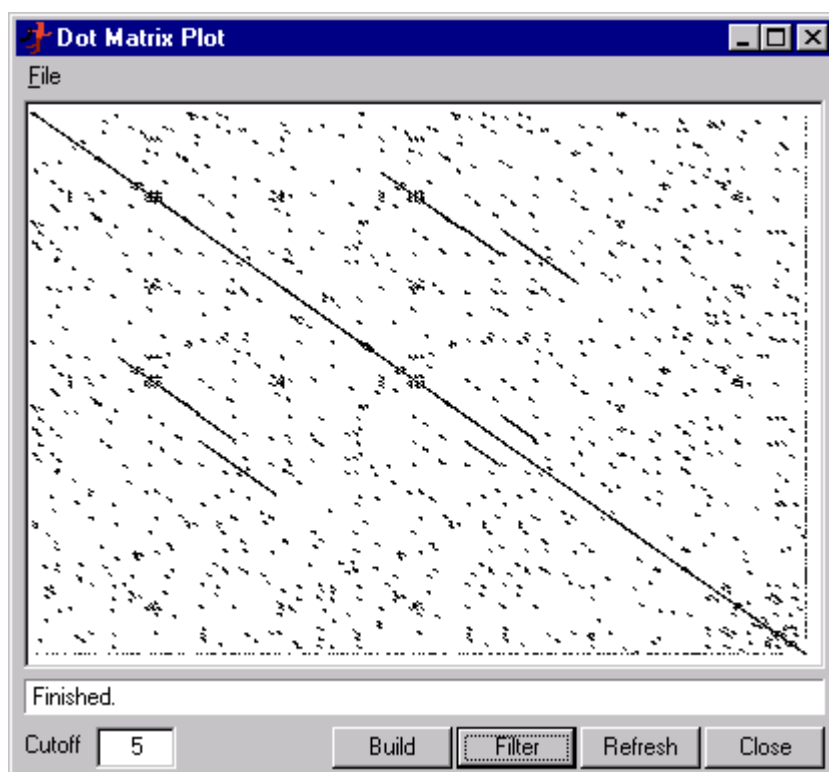


第二步，过虑完整的矩阵（例如移除不需要的点）。可以通过两种方法实现：1，重复的点击“过虑”，这样可以增加一个已存在的匹配的最小长度；2，在点击“过虑”之前，在文本域中输入一个值。如果用户选择一个小于当前值的中止值，有必要在它被过虑成新的更小的值之前重新建立一个完整的矩阵。这将会自动的完成。

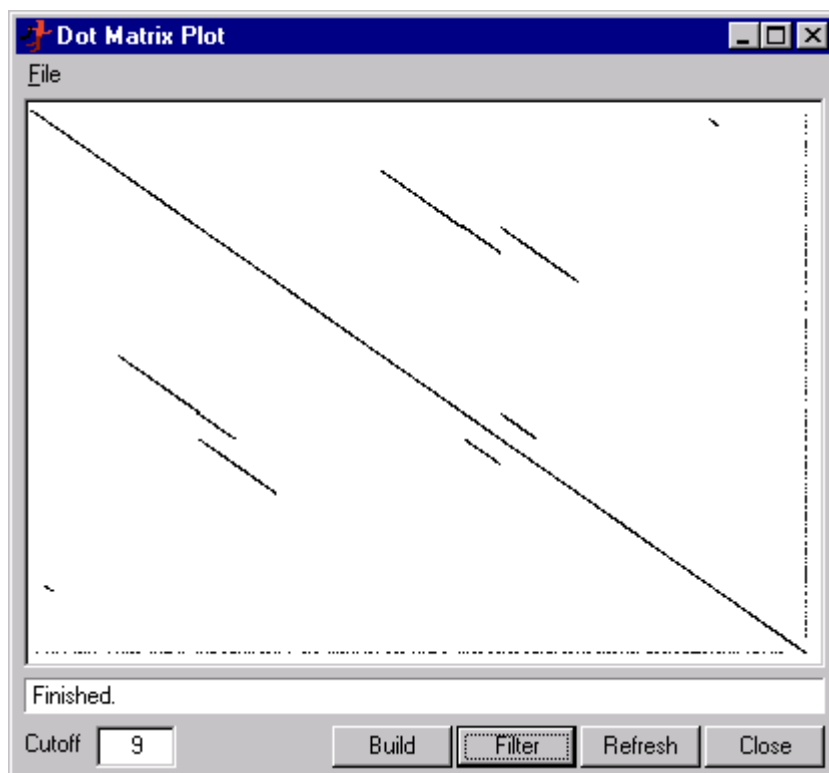
用最小值 4 进行过虑：



用最小值 5 进行过滤:



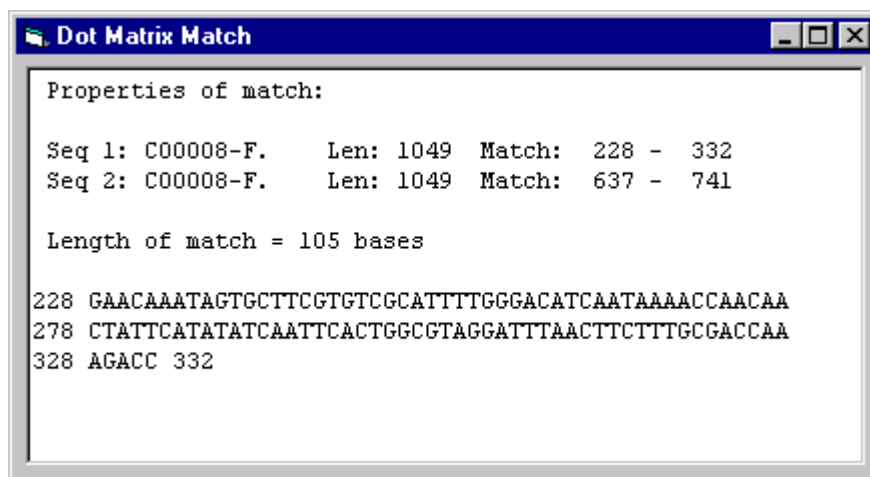
用最小值 9 进行过滤:

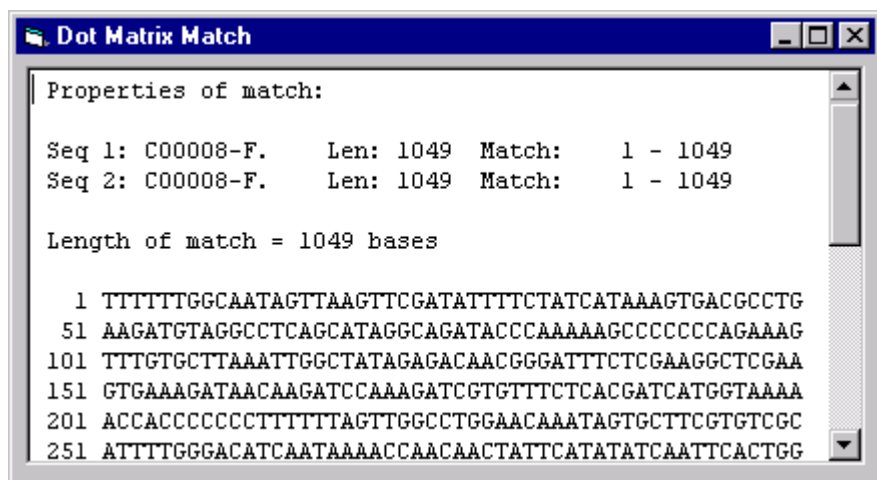
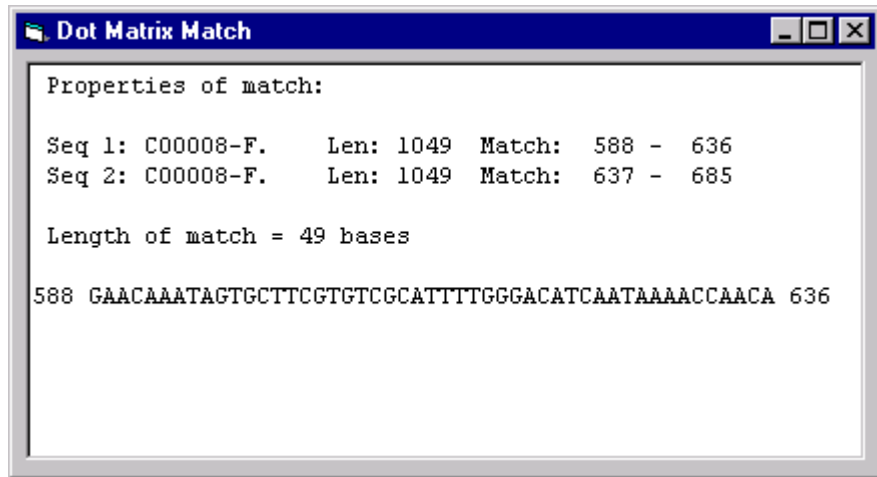


寻找和观察一个匹配:

当用指示器搜索点阵时按下左鼠标键, 展示在两个序列中的匹配的部分。释放左键并突然靠近最近的对角线(如果用户靠得足够近的话)以显示匹配的属性(对角线)。

Examples of retrieved matches derived from the dot plot shown above.



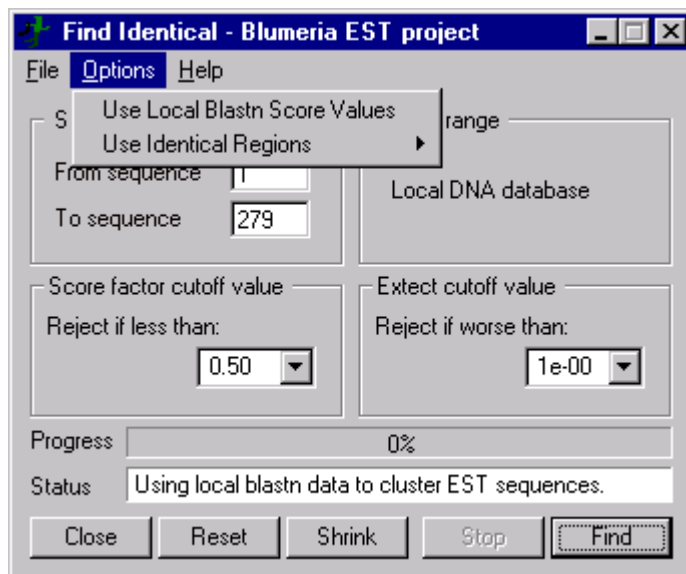


评述：

当前版本，这个功能只能找到完全匹配的区域。以后的版本将可以确定鉴别被一个或多个误配打断的对角线

4. 寻找相似性序列 blastn 数据：

使用本地 blastn 分数值评价配对，对所有当前方案中的序列进行比较。在执行比较前，在序列的标题中必须可获得本地 blastn 搜索的结果，例如每个序列必须包含一个本地的 blastn：在序列标题中的部分。



比较包括三步：1，从所有当前方案中的序列中提取自身分数值（和自身比较时得到的分数值）；2，使用自身分数值的分割法规范每个本地 blastn 配对的分数值；3，最后，规范化后的分数值与选定的分数中止限制进行比较同时低于中止值的配对被丢弃。

期望的中止选项允许用户依据配对的期望值排除配对。在大多数情况下，中止值应该设定为 1，例如定义一个充分大的值使得所有的配对都可用于分数值评估。选项同时还包括：允许用户排除那些能产生高于某个最小值的配对的序列对。

比较结果是一张序列对列表，其本地 blastn 配对优于中止值。使用这张表以生成不同类型的报告。例子见下：

Comments（评述） 略

This function will in most cases yield the same overall clustering of EST sequences as the Find Identical Regions routine and is considerable faster. There are, however, situations where the common identical regions method will yield a more reliable clustering: Long sequences with small overlaps from the same open reading frame may be omitted when blastn clustering is used but will, if they share just a short identical region be linked by the identical region method. On the other hand, sequences from different ORFs sharing a short identical region may erroneously be linked with the identical region method but not when local blastn data are used.

如何：

使用多序列功能以生成 FastA 源文件；

使用多序列功能以转化 FastA 源文件为一个本地核苷酸数据库；

针对核苷酸数据库执行本地 blastn 搜索；

使用 “Use Local Blastn Score Values” 选项运行寻找相似序列；

生成序列或克隆报告，见下：

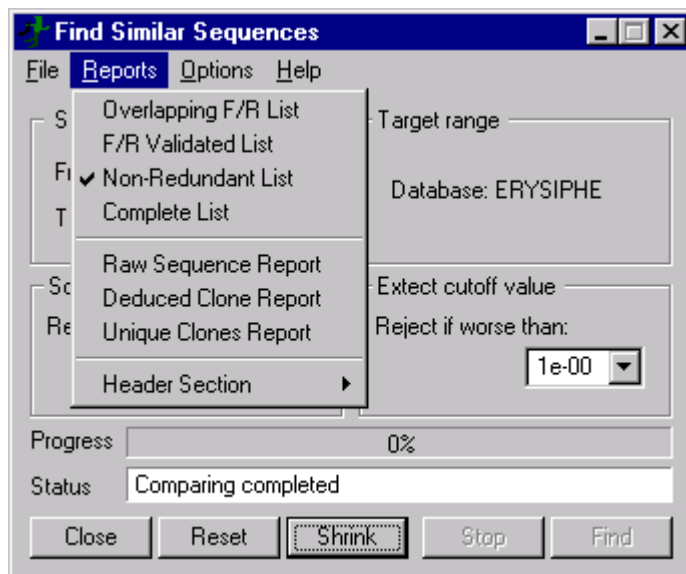
组类型：

重叠 F/R 序列—此列表包含来自相同克隆的 F 和 R 序列，其中这些克隆共享相同的区域或指定长度。对于双通过 EST 序列，提示 F 和 R 序列包含整个插入。伴随一致的序列命名，此选项可被用于显示来自同一模板的重叠序列。

R/F 验证列表—验证列表排除重叠的 F/R 序列，如上所说。由于插入太长以至于不能被前向和逆向序列或者那些特定的情况所包容如：对于一个给定的克隆，只可存在一个序列的情况。伴随一致的序列命名，此选项可被用于消除来自同一模板的重叠序列。

非多余序列列表—显示比对序列对的非多余列表。在列表中只包含一次每个比对序列对。

完整列表—按照字母显示比对序列对的完整列表。此列表包含同样的比对序列两次。暗示这些按照字母分类的列表将包含一个完整的序列列表，且这些序列和此方案中的其他序列共享一个相同的区域。



报告类型：

粗序列报告-对于输出列表中的每个比对序列对，这两个序列的名字都被与其他比对序列对的名字进行比较。万一一个比对序列对的名字和另外一个相同，将生成一个非多余名字链以产生一个粗序列报告，此报告是严格依照序列同一性。

一条链将包含所有的序列名字，且这些名字是鉴于相同配对的成员而被连接起来的。点击链的第一行，显示名字或包含于链中的序列的指定的标题行。

演绎克隆报告-演绎克隆报告是基于粗序列报告并且假定共享其文件名前 6 字符的序列指的是相同的克隆。演绎克隆报告的基本原则如下：

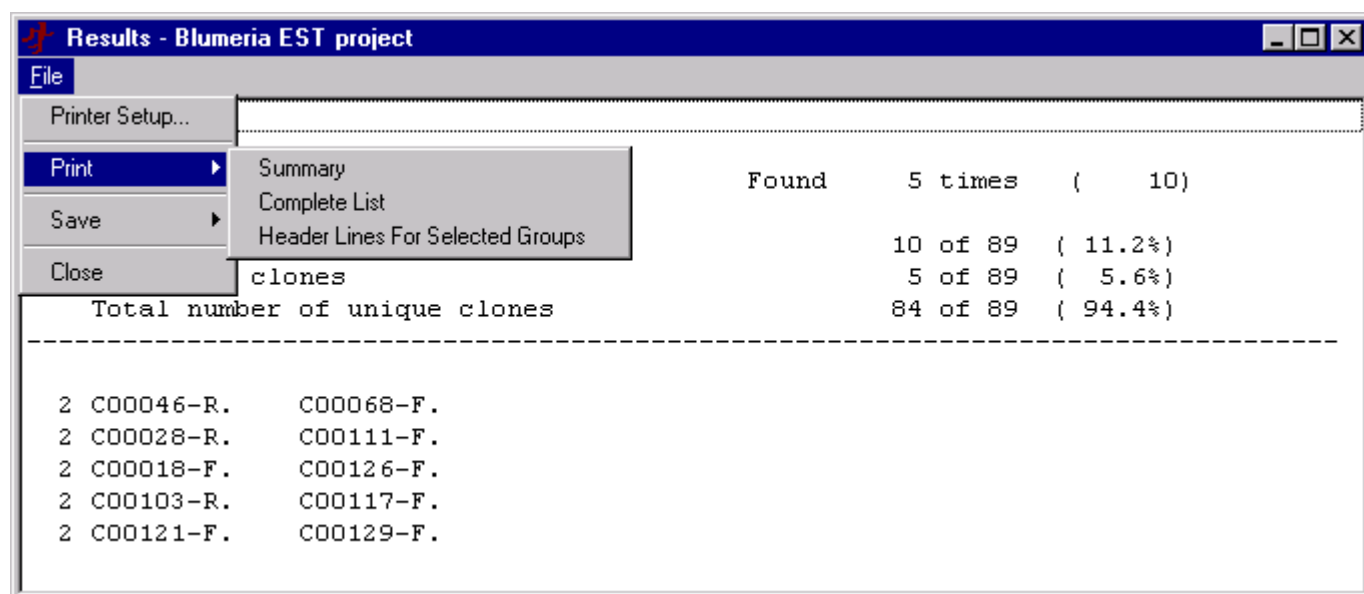
假定链 clone1-F - clone2-R 共享一个相同的序列区域；

假定链 clone2-F - clone3-R 共享一个相同的序列区域，此区域与上述的区域不同；

假定序列 clone2-R 和 clone2-F 来自同一克隆，四个序列可以被加入到一个新的链中：

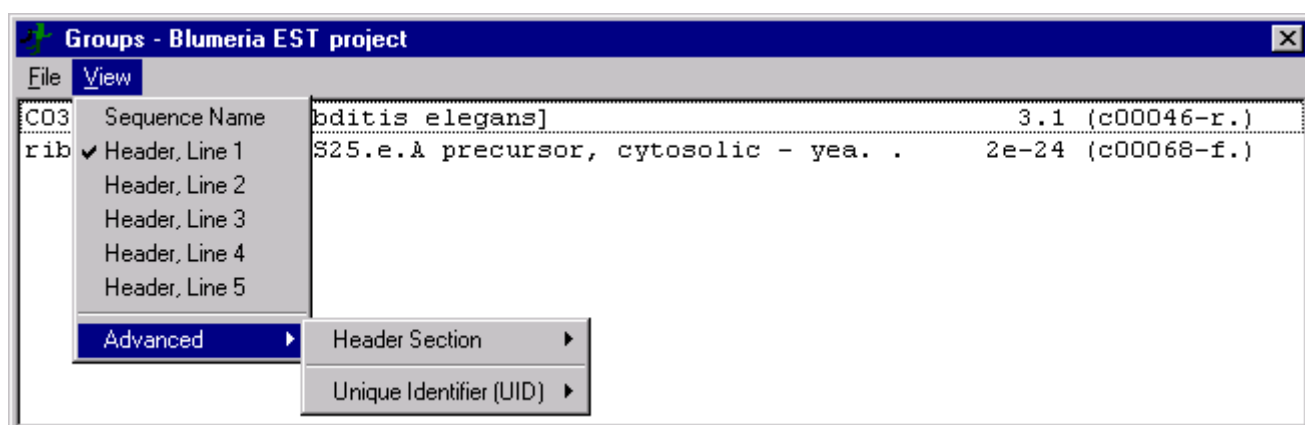
clone1-F - clone2-R - clone2-F - clone3-R；

加工那些属于新链的序列以移除来自同一克隆的完全相同的区域。通过分析序列标题和使用最小的信息标题否决序列来完成此项工作。



演绎的克隆报告可以以其在报告表中的展现形式或以一个选定的标题行列表形式被打印或保存下来，而不是以一个相同组中每个成员的克隆名形式。标题行格式与用于 View group form 中的格式相同。在保存和打印之前，首先选择格式，即打开“Identity group form”并选择一个“View option”。用于打印或保存的克隆组可在报告表中被选择，即当按下 CTRL 键时点击或者拖动鼠标。当选择好组之后，点击“*File/Print/Header Line For Selected Groups* 或者 *File/Save/Header Line For Selected Groups.*”

点击一个组的第一行，显示那个组中所有序列的注释。查看菜单，就像许多其他 DNAtools 表中的一样，允许用户定义序列标题的哪个部分用于组列表中的序列鉴定。



独特克隆报告一通过假定：或者直接的（就像在粗序列报告中的一样）或者通过第三个共享相同区域的序列加入方法（就像在演绎的克隆报告中的一样）的共享同一区域的两个序

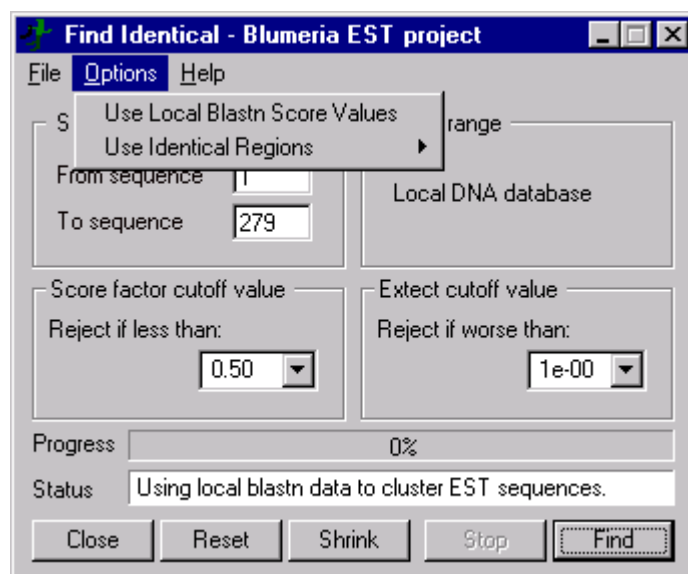
列都来自同一克隆，独特克隆报告是演绎克隆报告的进一步演化。独特克隆报告否决所有在链中的克隆，除了每个链的第一个克隆。这将为当前方案创建一个演绎的独特克隆列表。独特克隆报告被格式化为 microtite 板的 12*8wells 以帮助克隆阵列的生成。

关于序列名字的重要信息：

明显的，最后的两个报告类型要求文件名的前 6 个字符鉴定克隆。第 8 个字符是 F，则被定义为前向链；而 R 则被定义为逆向链。文件名的第 7 字符和文件扩展名的 3 个字符不被考虑

5. 寻找相似性序列 相同区域：

此项功能执行当前方案中的序列比对以寻找相同的区域。执行如下：对于每一个源序列，使用者选定的长度的视窗被用于搜索所有靶序列（单或双方向）。搜索视窗被移动一个 Step 长度同时不断重复搜索直到序列的末端。



如果找到了一个匹配，匹配被延长直到两个序列不同。搜索窗口的尺寸增加以满足延展的匹配长度同时靶序列的剩余部分也被搜索。如果没有找到更多的匹配，匹配长度被报告。视窗大小被重新设置同时针对接下来的靶序列重复搜索。程序不断重复直到所有的序列都被搜索过。

当搜索完成后，搜索结果的总结和搜索范围、参数等被显示在信息框中。匹配序列列表被显示在一个独立的列表中。

在关闭信息框后，双击匹配列表中的一个项目，自动的启动序列比对功能，重复两序列比对以核实该两个序列的匹配区域。

如果搜索以一个更大的搜索视窗进行重复搜索的话，只有那些在先前搜索中找到的匹配才会被搜索。降低搜索视窗的尺寸重新定义范围为当前方案中所有序列。

如果在当前方案中的较大数量序列中进行搜索且 step 长度较小，核实搜索将持续较长的时间。可点击 stop 终止搜索。若绿色条到达底端，搜索将继续。

可以延长一个已经存在的序列比对。查看“Extend identity search (*.ird file)”。

选项：

源和靶范围：搜索范围（源和靶序列）可以是当前方案中序列的任何组合。

最小匹配长度—从下拉列表框中选择一个新的值以改变搜索窗口大小（最小匹配长度），当然也可通过在文本域中输入一个值来实现。

敏感度—从下拉列表框中选择一个新的值以改变 step 长度。默认的 step 长度是视窗长度的一半但是可以被改为任何值—在 1 和视窗长度之间—列在下拉列表中。

1 个单位的 Step 长度将搜索所有可能的片断，然而等于片断长度的 step 值将以非重叠搜索窗口进行搜索。Step 长度的设置将很大的影响搜索的持续时间。

对于一个包含 2500500 碱基序列的方案，一个单位的 Step 长度搜索将花费许多小时，即使在一个快速的 PC，可能在晚上执行会更好。然而，此项工作可以在后台进行，因此可以在使用其他程序时进行比较，例如额外的 DNAtools 实例。

注意：用户可能会发现有不同的匹配数，这依赖于设置。只有 1 个单位的 step 值才会寻找所有相同的区域。一个更大的值可以减少搜索时间，但可能会错过一些相同的区域。

因为滑行窗口总是起始于序列的 5' 端并且当一个指定长度的区域不能再从序列中提取时终止，针对整个方案进行的沃森链搜索将产生一个稍微不同的结果。（相比于用互补的克里克链进行同样的搜索）。

文件菜单：

保存为一保存 R/F 核实过的，非多余列表，一个完整的列表或者克隆阵列。

打印：打印当前展示列表。

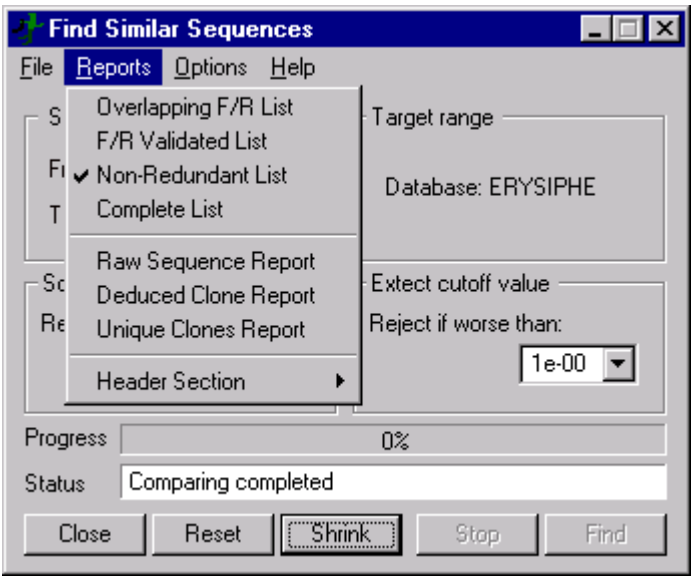
组类型：

重叠 F/R 序列—此列表包含来自相同克隆的 F 和 R 序列，其中这些克隆共享相同的区域或指定长度。对于双通过 EST 序列，提示 F 和 R 序列包含整个插入。伴随一致的序列命名，此选项可被用于显示来自同一模板的重叠序列。

R/F 验证列表—验证列表排除重叠的 F/R 序列，如上所说。由于插入太长以至于不能被前向和逆向序列或者那些特定的情况所包容如：对于一个给定的克隆，只可存在一个序列的情况。伴随一致的序列命名，此选项可被用于消除来自同一模板的重叠序列。

非多余序列列表—显示比对序列对的非多余列表。在列表中只包含一次每个比对序列对。

完整列表—按照字母显示比对序列对的完整列表。此列表包含同样的比对序列两次。暗示这些按照字母分类的列表将包含一个完整的序列列表，且这些序列和此方案中的其他序列共享一个相同的区域。



报告类型：

粗序列报告：

粗序列报告-对于输出列表中的每个比对序列对，这两个序列的名字都被与其他比对序列对的名字进行比较。万一一个比对序列对的名字和另外一个相同，将生成一个非多余名字链以产生一个粗序列报告，此报告是严格依照序列同一性。

一条链将包含所有的序列名字，且这些名字是鉴于相同配对的成员而被连接起来的。点击链的第一行，显示名字或包含于链中的序列的指定的标题行。

演绎克隆报告-演绎克隆报告是基于粗序列报告并且假定共享其文件名前 6 字符的序列指的是相同的克隆。演绎克隆报告的基本原则如下：

假定链 clone1-F - clone2-R 共享一个相同的序列区域；

假定链 clone2-F - clone3-R 共享一个相同的序列区域，此区域与上述的区域不同；

假定序列 clone2-R 和 clone2-F 来自同一克隆，四个序列可以被加入到一个新的链中：

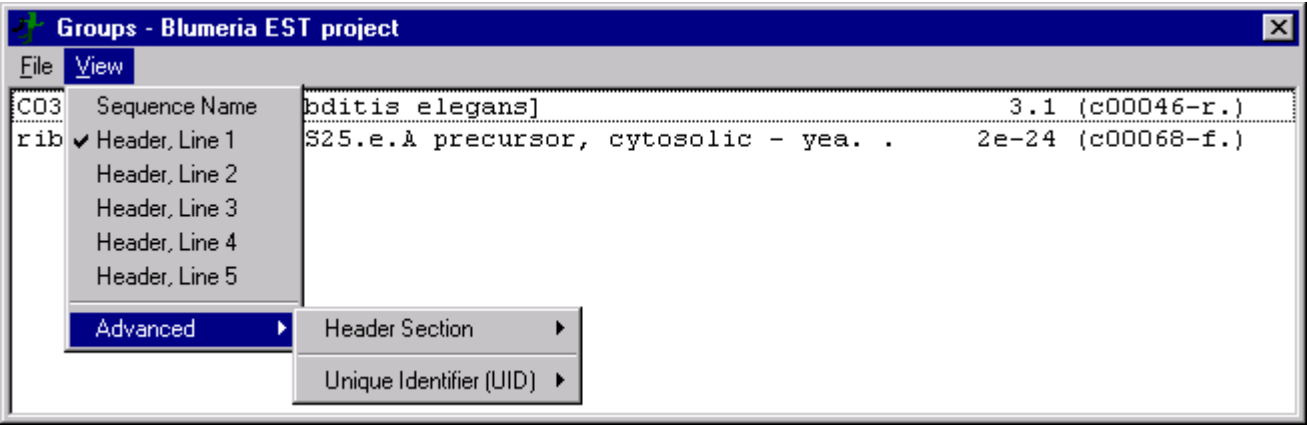
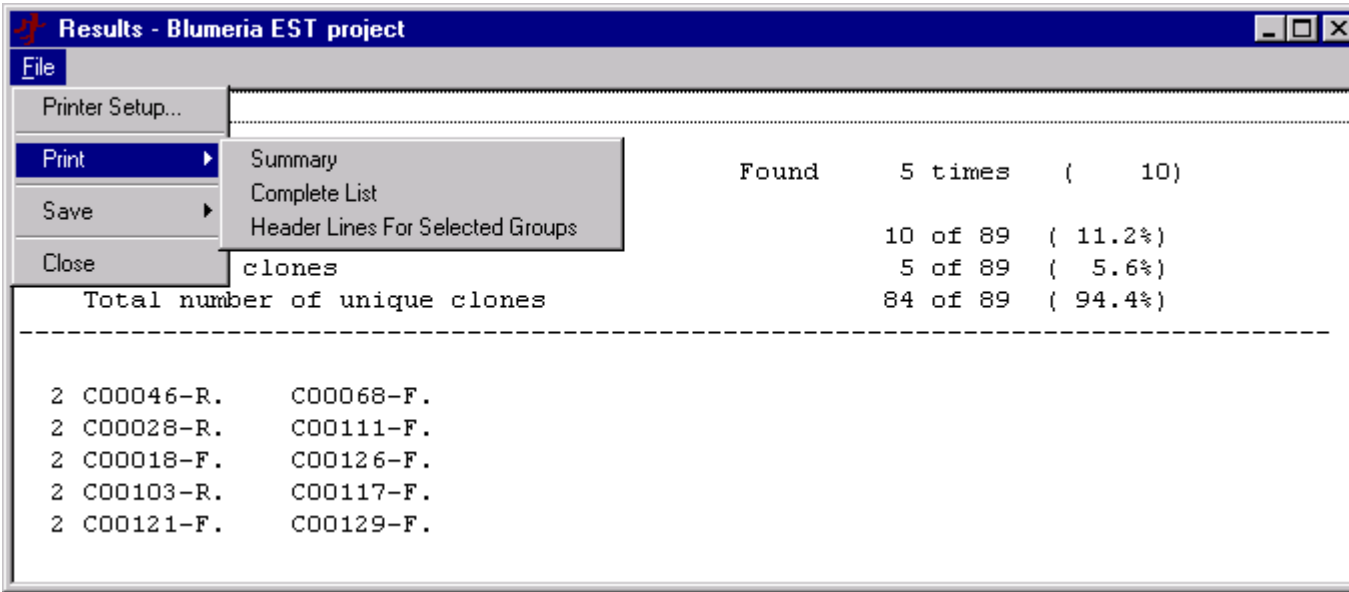
clone1-F - clone2-R - clone2-F - clone3-R；

加工那些属于新链的序列以移除来自同一克隆的完全相同的区域。通过分析序列标题和使用最小的信息标题否决序列来完成此项工作；

在演绎的克隆报告中，这两个第一链被一个单个链所取代：clone1-F - clone2-F - clone3-R（如果克隆 2-R 有最小信息标题）

演绎的克隆报告可以以其在报告表中的展现形式或以一个选定的标题行列表形式被打印或保存下来，而不是以一个相同组中每个成员的克隆名形式。标题行格式与用于 View group form 中的格式相同。在保存和打印之前，首先选择格式，即打开“Identity group form”并选择一个“View option”。用于打印或保存的克隆组可在报告表中被选择，即当按下 CTRL 键时点击或者拖动鼠标。当选择好组之后，点击“*File/Print/Header Line For Selected Groups* 或者 *File/Save/Header Line For Selected Groups.*”

点击一个组的第一行，显示那个组中所有序列的注释。查看菜单，就像许多其他 DNAtools 表中的一样，允许用户定义序列标题的哪个部分用于组列表中的序列鉴定。



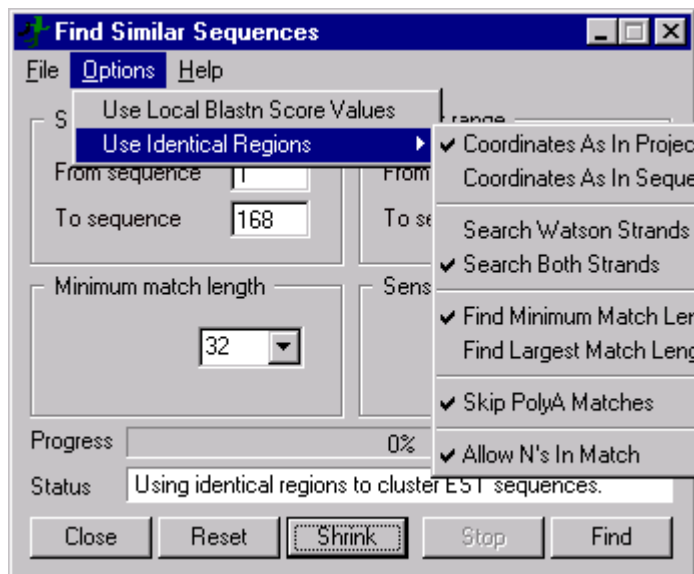
独特克隆报告一通过假定：或者直接的（就像在粗序列报告中的一样）或者通过第三个共享相同区域的序列加入方法（就像在演绎的克隆报告中的一样）的共享同一区域的两个序列都来自同一克隆，独特克隆报告是演绎克隆报告的进一步演化。独特克隆报告否决所有在链中的克隆，除了每个链的第一个克隆。这将为当前方案创建一个演绎的独特克隆列表。独特克隆报告被格式化为 microtite 板的 12*8wells 以帮助克隆阵列的生成。

明显的，最后的两个报告类型要求文件名的前 6 个字符鉴定克隆。第 8 个字符是 F，则被定义为前向链；而 R 则被定义为逆向链。文件名的第 7 字符和文件扩展名的 3 个字符不被考虑

Coordinates

在当前序列中—为当前方案中的序列的起始部分外加相配之物给相同的区域。如果在转换序列为其互补序列时找到了一个匹配，在结果列表中匹配将被标记为 C。

在匹配序列中—为源序列（在这些序列中找到了匹配）的起始部分外加相配之物给相同的区域。为了在方案中的序列中定位匹配，在克里克链上找到的匹配（被 C 所标记的）必须先被转换为其互补序列。



链：

沃森链—用源序列的沃森链进行搜索。

两条链—搜索源序列的两条链。

区域：

找到最大相同区域—此选项最耗时间，除非用户真的期望看看这些最长的相同区域，否则还是不要使用它。

只找视窗长度—紧紧寻找但是并不延展最初的相同区域。选项是默认的并且是优先使用的以生成序列相同区列表。双击列表中的一个项目激活比较功能，这将产生一个完整的相同区域列表。

关于序列名字的重要信息：

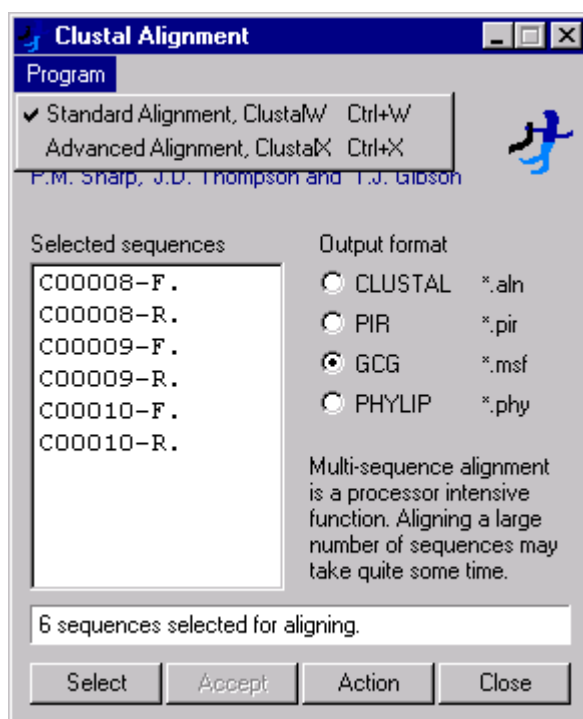
明显的，最后的两个报告类型要求文件名的前 6 个字符鉴定克隆。第 8 个字符是 F，则被定义为前向链；而 R 则被定义为逆向链。文件名的第 7 字符和文件扩展名的 3 个字符不被考虑。

6. 使用 clustal 进行序列比对：

此项功能使用 Clustal 程序执行真正的配对。ClustalW 程序被完全整合到 DNAtools 中并且不太可能改变默认的参数，但是 ClustalX 却是卓越的视窗程序。不像 GeneDoc，可以从作者的主页直接下载它，而两个 clustal 测序只可以从 DNAtools 下载页面进行下载。

Clustal W：

执行比对按照步骤一直往下即可，按照此页底部的教程



输出格式：

clustal 默认*.aln-默认的 clustal W 格式含方案名字和日期;

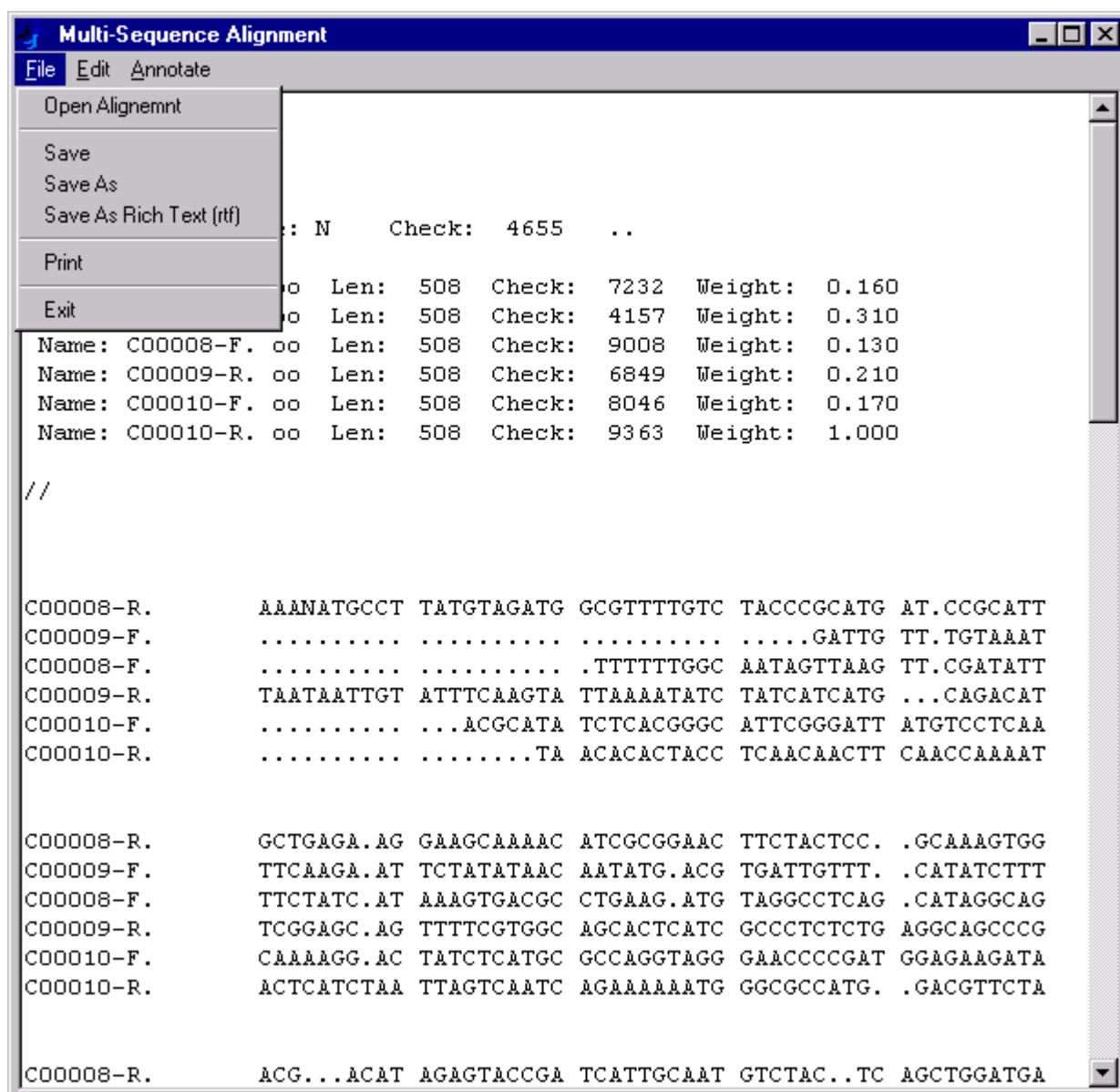
PIR, *.pir-配对的序列作为分离的序列被保存为 FastA 多序列格式;

GCG, *.msf-保存为此种格式的多序列配对可以被输入进 GeneDoc 中, 一个高级配对编辑器, 看如下;

Phylip. *.phy -

从一个 clustal W 配对中输出:

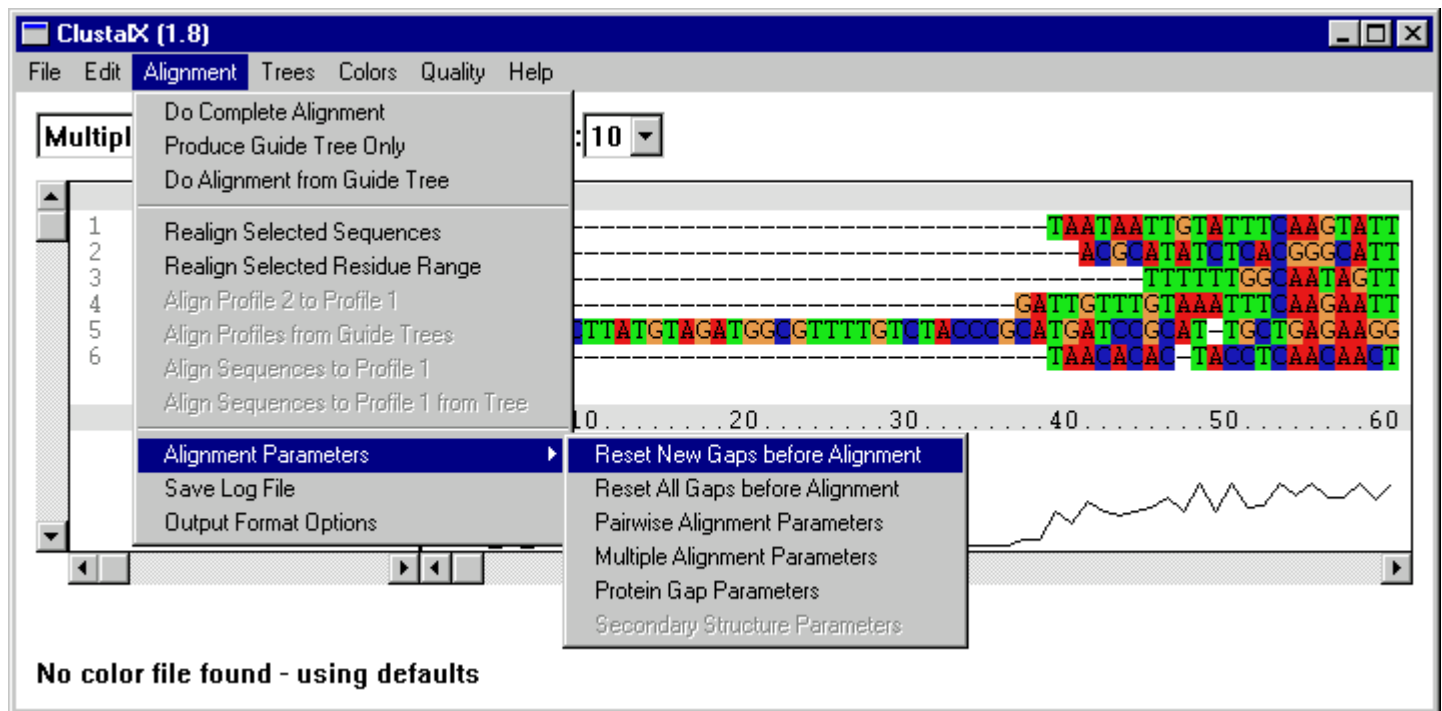
来自配对的输出结果被显示在一个简单的文本编辑器中, 但只含有限的选项由于注释。



如果 GeneDoc 已经被正确的安装在用户的 PC 上且被整合在 DNAtools 的启动菜单中，并且 clustal W 的输出格式被定义为*.msf，那么 GeneDoc 会出现在输出表的菜单行（在输出表的上方不可见）。点击“GeneDoc”会自动的打开 GeneDoc 程序和多序列配对 (*.msf 格式) 用于进一步的编辑和注释。

Clustal X 高级配对：

Clustal X 是一卓越的视窗程序。在选择需要比对的序列之后，从 DNAtools 中选择此选项会导致序列被转载入 ClustalX 中。比对的剩余部分，包括为比对设置参数，可以在 Clustal X 程序中进行。查看下载页寻找进一步的信息。



如果用户希望进一步在 GeneDoc 中编辑 Clustal X 产生的比对，比对必须被保存为*.msf 格式。因为在 clustal X 和 DNAtools 中的 GeneDoc 之间没有直接的联系，输出的*.msf 文件必须经 GeneDoc 文件菜单被装载到 GeneDoc 中。

如何。。。。。

装载序列(DNA or protein)入一个方案中；

打开比对表，*Search/Align Sequences*；

从文件列表中选择需要比对的序列；（当按下 CTRL 键时点击左键）

接受选择；

选择输出格式；

点击 *Action* 等候；

当 *View Alignment* 命令按钮出现时，点击它查看比对。

GeneDoc:

为了更好的利用 GeneDoc 程序，用户必须首先下载并安装它。然后，通过使用 DNAtools 的 *Preferences/General* 菜单在主编辑器的启动菜单中生成一个入口。这将使得 DNAtools 可以从多序列比对的结果表菜单中获得 GeneDoc。注意：只有当输出结果是 GCG 格式时，GeneDoc 菜单选项才可以看见。

作者（略）

Author – Multiple sequence alignment editor & Shading Utility, Version 2.5.002. , Copyright 1999 by Karl Nicholas.

从 GeneDoc 帮助文件中提取：

GeneDoc 作为一个常规的视窗程序被安装入 GeneDoc 程序组中。在启动 GeneDoc 之后，使用选项 “file open and read in a MSF (Multiple Sequence File) file” — 在 MSF 多序列文件中打开和读取文件。GeneDoc 在 MSF 的评述部分为这些文件保存设置信息，因此如果 GeneDoc 保存了这个文件，它将以相同的设置被重新打开。

读取/输入数据（帮助索引）— 用户可以输入非 MSF 文件。使用 “File/New menu” 菜单，然后选择 “File/Import”。输入对话框允许从剪贴板、磁盘文件或手动输入。Clustal, fasta 和其他一些类型可以以这种方式被读取。

GeneDoc 网页浏览器（帮助文档）— GeneDoc 可以在其他程序如网页浏览器或数据库程序中运行。

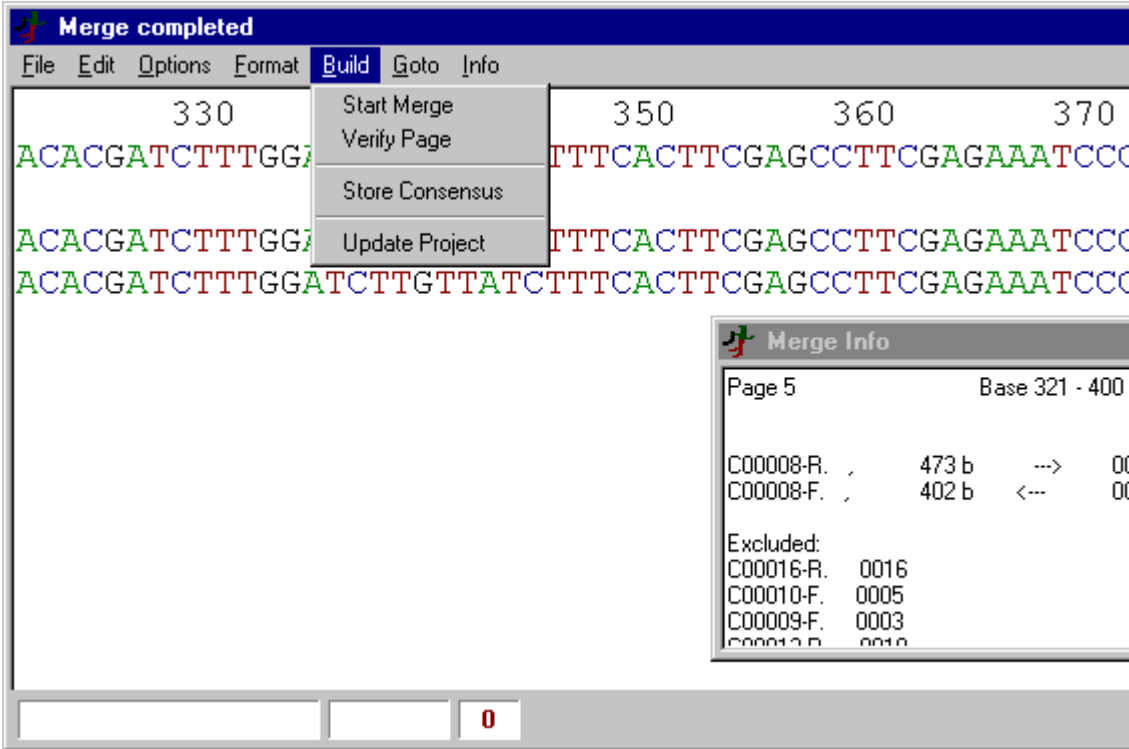
DDE 支持（帮助文档）— GeneDoc 提供最小的 DDE 支持，文件打开和打印。这将允许用户通过双击或右击 MSF 文件图标来自动的启动视窗打开文件或打印文件。

GeneDoc 可以免费获得，无需任何保证。

7. 连接编辑：

这项功能旨在帮助用户建造和编辑重叠的核苷酸序列。

连接编辑菜单



命令按钮用于选择，核实和合并序列：



队列寻求子序列所有匹配的最大相同区域。然后使用这些信息展示合并的，排列的相同区域。在队列加工过程中，并没有考虑队列区域上游和下游的序列，提示：只有在进行以下操作即：“在合并子序列之前，编辑这些亚序列以去除低质量区域”时这项功能方才有用。同时要求亚序列属于同一连接，如来自同样的基因/开放读码框/核苷酸片断。

Page up 和 page down 移动序列队列到队列的下或前一页；

按下 CTRL 显示一个小的文本域用于输入序列序号，按下回车键进入到选择的基数；

在合并中使用箭头键四周移动，序列名、长度和当前鼠标位点的基数被显示在合并下的两个文本域中；

所有包含在当前页中的序列和其在合并的定位被显示在不同的表中，其中使用从合并中排除出来的序列的名字和箭头提示这些定位；

在合并编辑中那些被升级后的编辑序列必须被保存为一个常规的方案，如果用户希望保留这些改变的话；

合并编辑器接受至少 200,000 碱基的序列(on a Pentium 266 MHz, 98 Mb RAM).

如何。。。。。

装载那些用户希望合并进入 DNA 序列方案中的序列；

点击选择以显示文件列表；

高亮显示将被合并的序列（在点击时按下 CTRL 键）；

点击 Accept 以包含在合并中的高亮序列；

点击 Merge 执行；

点击 Update project 以使用进入合并中的变化升级方案；

点击 *Store Consensus* 将同意附加到当前方案中；

从主菜单中选择每行的基数；

点击 *File/Print Merge*. 打印合并。

评述：略

8. 运行 DOS 程序：

这里提及 Clustal W 和双 blast 程序。因为长文件名和目录名是不被 DOS 程序支持的，DNAtools 将 DOS 相关的操作转移到一个分离的目录，**DT5_TEMP** 定位在 **Windows/Winnt** 下的一个子目录。当 DNAtools 启动 Clustal 和 Blast 功能时，**exe** 文件则被拷贝到这个子目录同时 DNAtools 的主目录下的文件和 Blast 数据子目录被删除。在运行过程中，DOS 程序产生的输入和输出文件被定位到这个目录。接着 DNAtools 找回这些结果文件用于进一步的加工。用户不用担心 DT5_TEMP 目录，因为 DNAtools 会自动的移除使用过的文件。

Chapter5: DNAtools-SAGE functions

1. SAGE 标签和双标签抽提：

这张表包含三个程序用于从不同类型的序列文件中提取 SAGE 标签。从 GenBank 文件，DNAtools 方案中的序列和方案中的标签序列文件中提取 SAGE 标签。除此之外，有另一程序用于生成 SAGEmap 文件。

从 GenBank 中提取：

这项功能从格式化和修剪过的 GenBank 文件（扩展名为*.tgf 或*.ngf）中提取 SAGE 标签。按照以下准则进行提取：在 2500 3' 最多参数中搜索序列（万一序列较长）直到最接近 3' 端的锚定位点被定位。然后，提取这个指定长度的 SAGE 标签。当所有的序列被提取后，标签列表被加工（包括在标签的第一次出现时复制标签，参考以下的例子）



用户可以限制提取序列在一个范围内如只含 polyA 尾的序列或者提取所有序列。含 5' polyT 区域的序列总是被转为其互补序列，尤其是当 5' T 超过用户定义的值时。

有可能生成一个 SAGEmap 可靠图谱文件 (*. smf) 用于使用标签抽提来同时鉴定标签。

注意：使用者定义标签长度不能包含锚定序列的长度，通常是 4 个碱基。SAGE 标签文件的扩展名为 *. stf (SAGE 标签文件)。

注意：用户无需首先装载序列入方案即可自己建造 GenBank 文件。只需收集序列到相同的目录并且使用多序列功能以建造 GenBank 多序列文件。这个特征允许用户建造非常到的标签文件而无需装载序列入方案。

从当前方案中提取：

这项功能与从 GenBank 中提取的工作方法相同，除了待提取的序列得事先被装载入 DNAtools。



除了标签文件 (*.stf), 此项功能产生两个更多的文件, 一个是 (*.lst) 包含那些不生成 SAGE 标签的序列列表; 另一个是 (*.psg) 可以被用于打开一个包含所有可生成 SAGE 标签序列的方案。

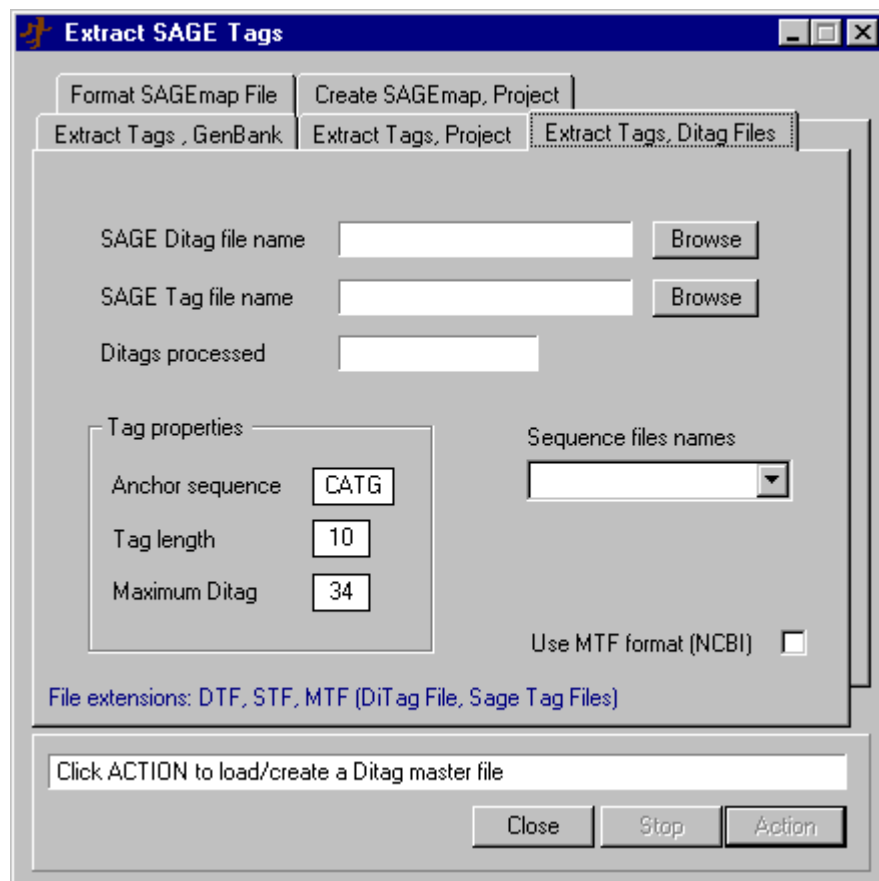
用户可以限制提取序列在一个范围内如只含 polyA 尾的序列或者提取所有序列。含 5' polyT 区域的序列总是被转为其互补序列, 尤其是当 5' T 超过用户定义的值时。

有可能生成一个 SAGEmap 可靠图谱文件 (*.smf) 用于使用标签抽提来同时鉴定标签。由于图谱信息可从序列标题中寻回, 只有当序列被注释时该功能方才起作用。查看电子邮件 Blast 搜索。

注意: 用户定义的标签长度不包括锚定序列的长度, 通常是 4 碱基。SAGE 标签文件的扩展名为 *.stf。

从双标记序列中提取：

在提取前，双标记序列必须事先被装载到 DNAtools 方案中，并且必须生成一个主双标记文件(*.dtf) 或者已经打开一个已经存在双标记文件。用户得输入一个期望的最大双标记长度 ($2 \times \text{锚定长度} + 2 \times \text{标记长度} + \text{少许以允许 II 型酶的变化}$) 同时选择一个文件名用于 SAGE 标记文件。



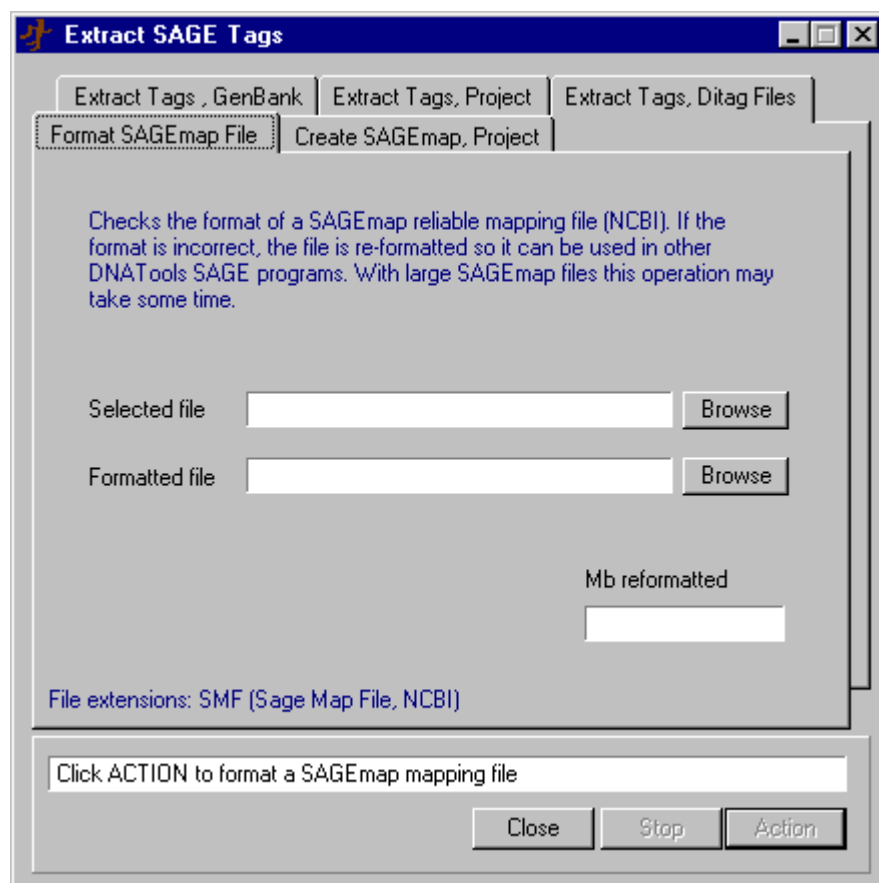
在方案中每个序列的 5' 端开始，两个锚定序列划界的区域被隔离开。如果两个全长的标记可以被提取且如果这个双标记比指定的最大长度短或相等的话，则审查双标记的长度。如果审查通过，该双标记序列则和主表中事先提取的双标记进行比较。如果这个标记已经存在则被拒绝掉。如果这个审查也通过的话，左右标记则被隔离开。下游标记则被转化为其互补序列。若标记不含 N，则他们会被包含于粗标记列表中。

当所有方案中的双标记文件被提取后，将会生成一个单一或加工后的 SAGE 标记列表。每个单一标记的拷贝数被记录下来并且加工后的标记列表被保存为*.stf 文件。依据标记数进行

分类*.stf 文件。升级后的主双标记列表也被保存为*.dtf 文件，这些文件包含所有的单一双标记序列和他们的长度。

SAGE 提取功能所生成文件的格式：

STF—SAGE 标记文件：包含一个文本标题，该标题含以下例子中提示的信息；该标记文件还含一个标记列表，此列表中的每一行都包含标记序列（不含锚定序列），找到标记的次数和标记的起始信息（用于提取于方案中序列标记的克隆名或用于非 DNATools 生成的 GenBank 多序列文件的索取号）。起始域的最大长度是 12 字符。过多的字符将被截断。对于提取自双标记序列的标记，起始信息是不被记录的。

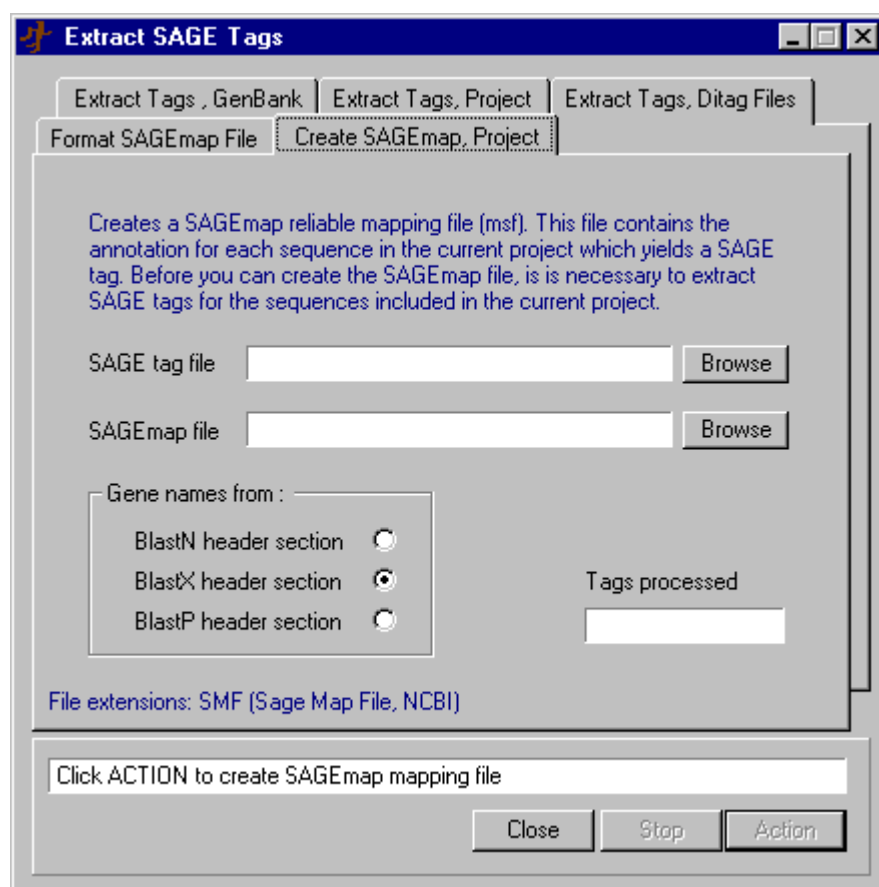


标题和文件列表被标准的 DNATools 分隔物分开：CR+LF + .. + CR+LC. 每个记录(= line)中的域(= words)被 tabs, chr(9) 分开。记录则被 CR + LF 分开。

SAGE 提取功能生成的文件类型：

MTF - Minimal sage Tag File (NCBI):最小 SAGE 标记文件：是一纯文本文件，包含标签序列（无锚定序列）和找到标签的次数。标签按照字母进行分类。每个记录(= line)中的域(= words)被制表符分开。同样的格式用于 NCBI 的可下载 SAGE 数据文件。

SMF - Sage Mapping File (NCBI): SAGE 图谱文件:是一纯文本文件，包含标签序列，克隆名/索取号和来自基因库多序列文件或者来自方案中序列的标题的基因名。每个记录(= line)中的域(= words)被制表符分离开。同样的格式用于 NCBI 的可下载 SAGEmap 可靠图谱文件。



DTF - sage DiTag File:SAGE 双标记文件：含标题，此标题包含以下例子中提示的信息和一个包含双标记序列和长度的双标记列表。标题和文件列表被标准的 DNAtools 分离物分开：CR+LF + .. + CR+LC. 每个记录(= line)中的域(= words)被 tabs, chr(9) 分开。记录则被 CR + LF 分开。

LST - LiST file:一个纯文本文件，要用文本编辑器才能查看。

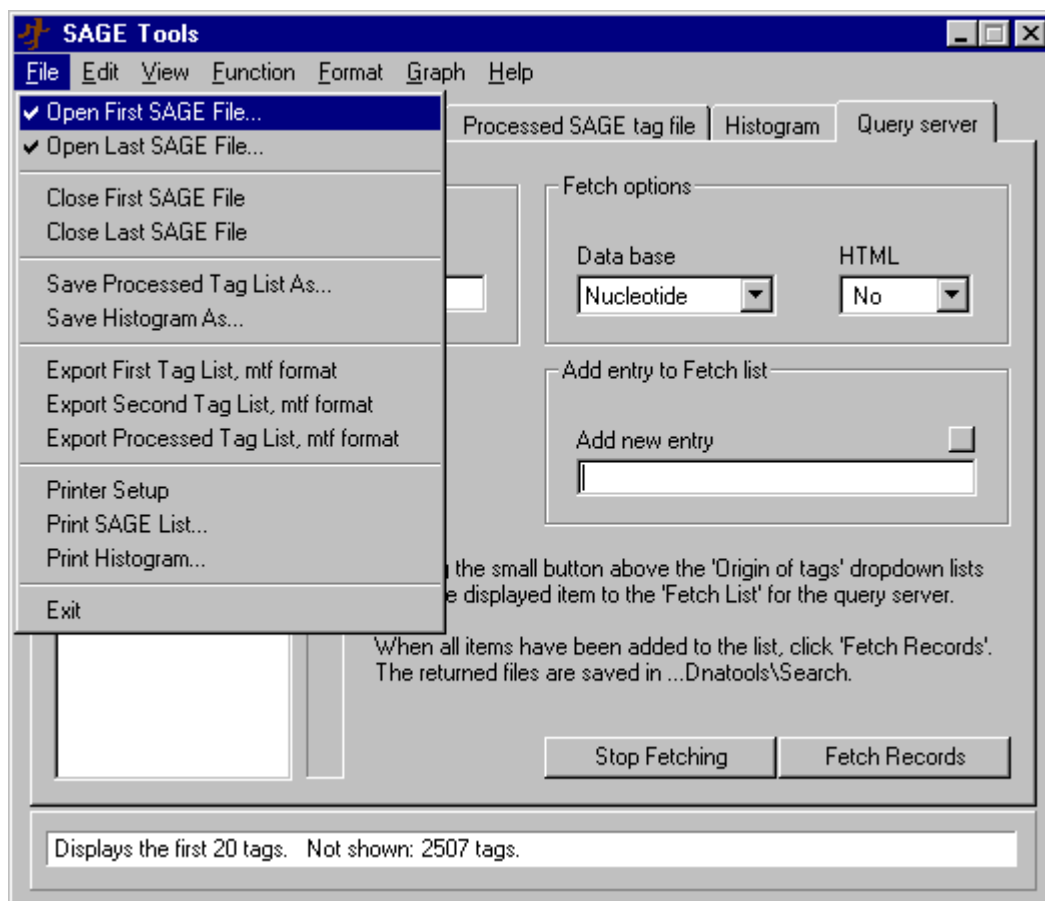
PSG - Project Sub-Group file 方案亚组文件：含所有产生 SAGE 标记的文件/序列的全路径（文件从那里被装载）。查看 Open Existing Project 寻找进一步的信息。

An example of a STF file:

PROJECT	Tags extracted from GENBANK. FGF		
FILE NAME	Tags from genbank. stf		
DATE	12-02-98	23:09:51	
ANCHOR	CATG		
LENGTH	10		
NUMBER	516		
DUPLICATES	608		
..			
GGATTCATGG	tab	47	tab ;X234765 ;D244765 ;H986556
ACGATTCGTT	tab	43	tab ;R223545 ;A445678

2. SAGE 轮廓工具：

这张表含一个功能集，用于修改和比较 SAGE（基因表达的序列分析）标记文件。



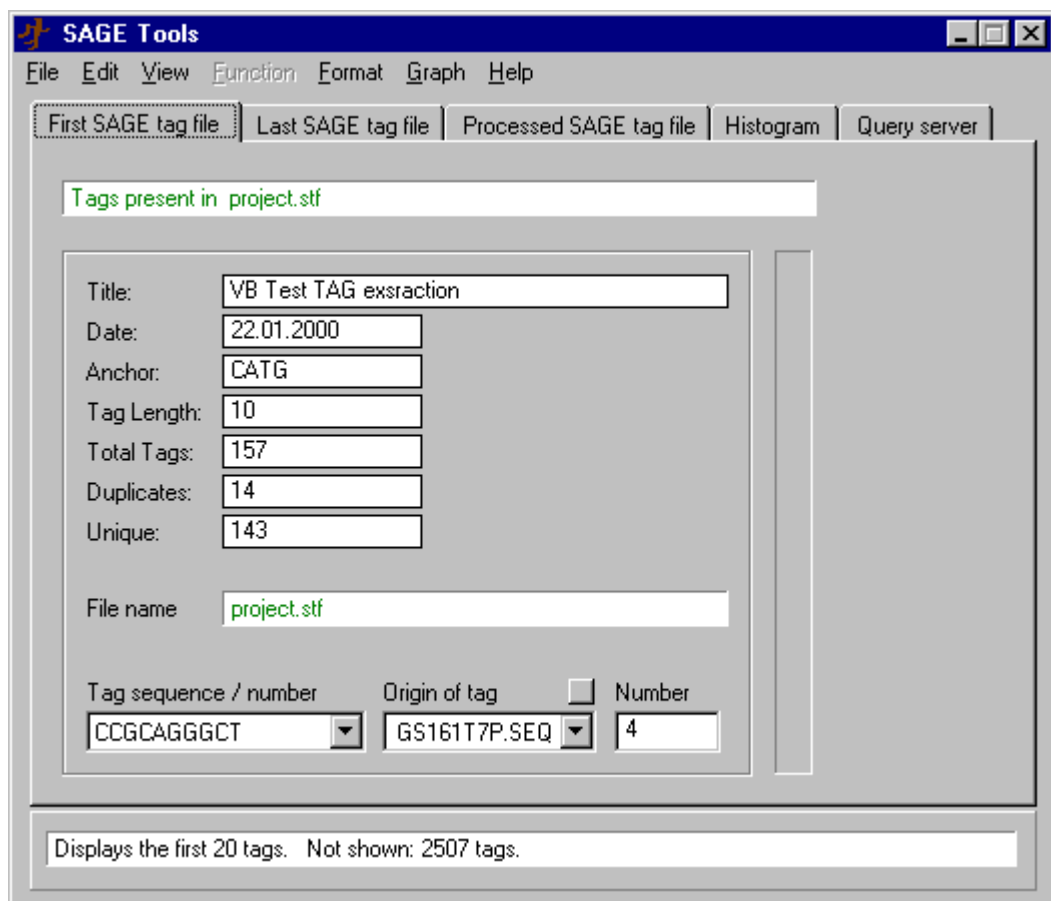
基因表达的 SAGE 分析是一个很有效的用于研究基因表达的方法，尤其是如果用户对细胞循环中两个不同阶段的表达模式感兴趣的话。当两个阶段的 SAGE 标记文件都可获得的话，这里讲述的功能将允许用户可以找到某个阶段或两个阶段表达的基因。询问服务器使得寻回数据库入口去取相关的基因变得很容易。

加工后的文件，如标记数被频率或百分率所取代，不能用于进一步的比较但是只是用于搜索装载于方案中的序列。加工后的标记文件的扩展名为*.pst。

文件：

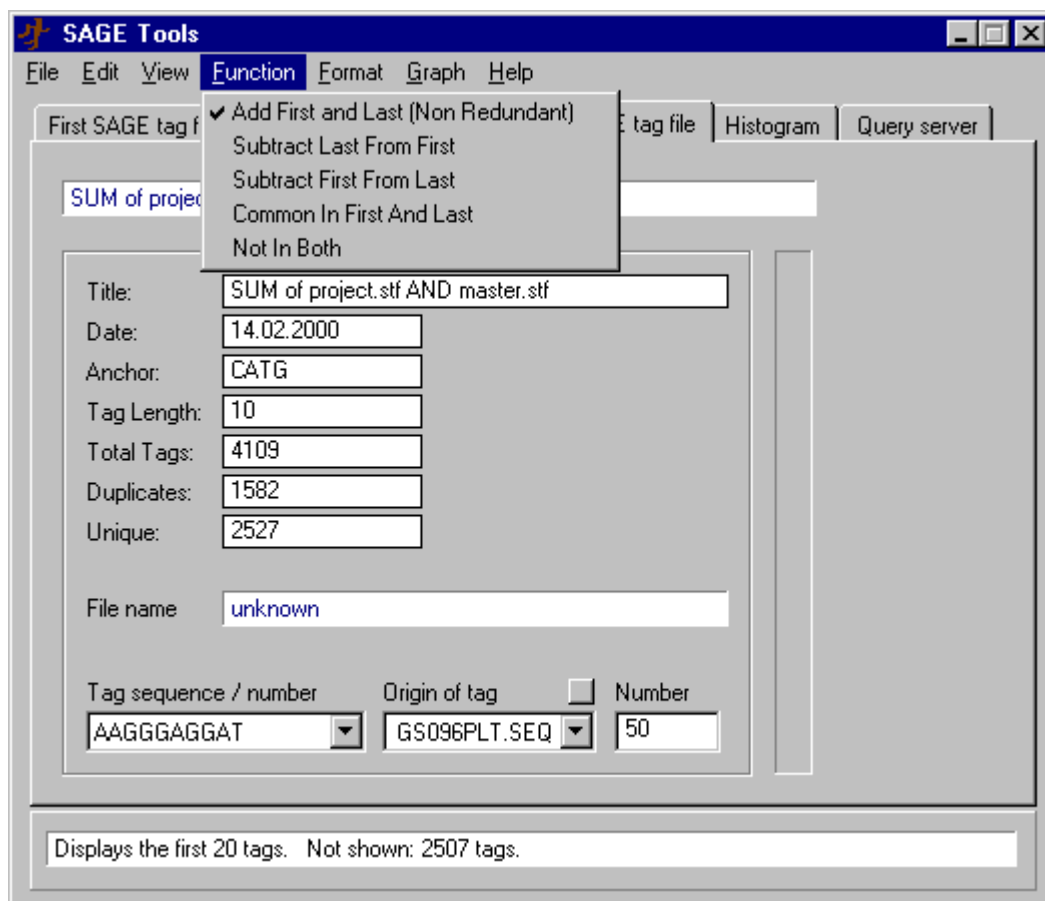
文件菜单含常规项目。SAGE 标记文件可以以默认的 DNAtools 格式 (*.stf) 或 NCBI 最小格式 (*.mtf) 被装载/输入和保存/输出。

所有 SAGE 工具中的功能要求装载两个 SAGE 文件且这两个文件是可相容的，如锚定和标记长度是相同的。加工后的文件可以以默认的 DNAtools 格式 (*.stf) 或 NCBI 最小格式 (*.mtf) 被保存/输出。



功能:

在菜单项目下，有五个功能可用于加工 SAGE 标签文件对。



如果一个功能没有返回一个结果，将显示出来一个错误信息“新的标记列表是空的”“The new tag list is empty”。如果用户尝试减少一个双标记列表并且他们中的一个的所有入口被包含在另一个中—或者若用户寻找两个文件的共同标记。

计算：

计算功能分析一个完整的标记文件。提取包含在文件中的标记的总数并计算每个与文件中标记总数有关的标记的频率或百分率。结果显示在“加工后的 SAGE 标记文件制表 Processed SAGE tag file tab”下并且最高频率被展示在柱状图中。嵌在柱状图中的值以“*”标记。

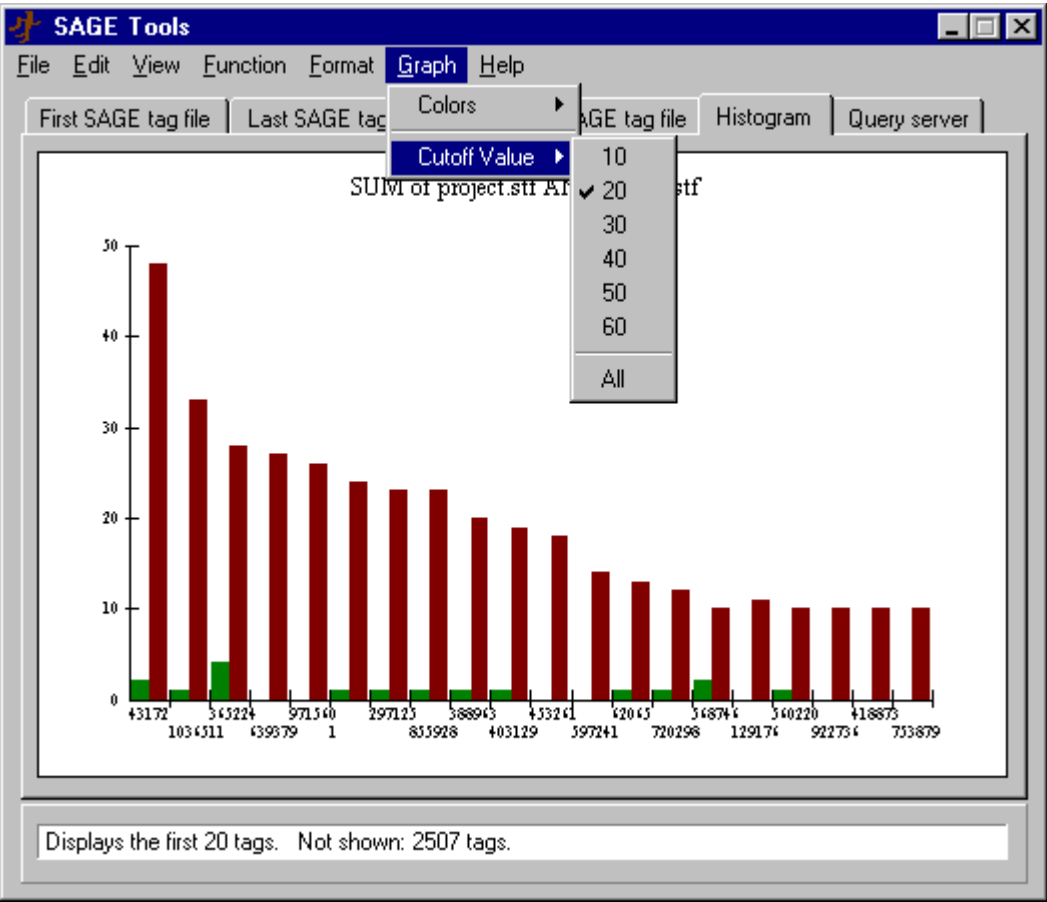
若两个 SAGE 文件的共同标记已被提取“Common in first and last function”，对于这两个初始文件的每个文件的分布，其频率或百分率被分别计算。结果显示之阿柱状图中，并以不同的颜色显示。

查看选项：

是通过序列查看 SAGE 标记还是通过标记 ID 来查看 SAGE 标记。

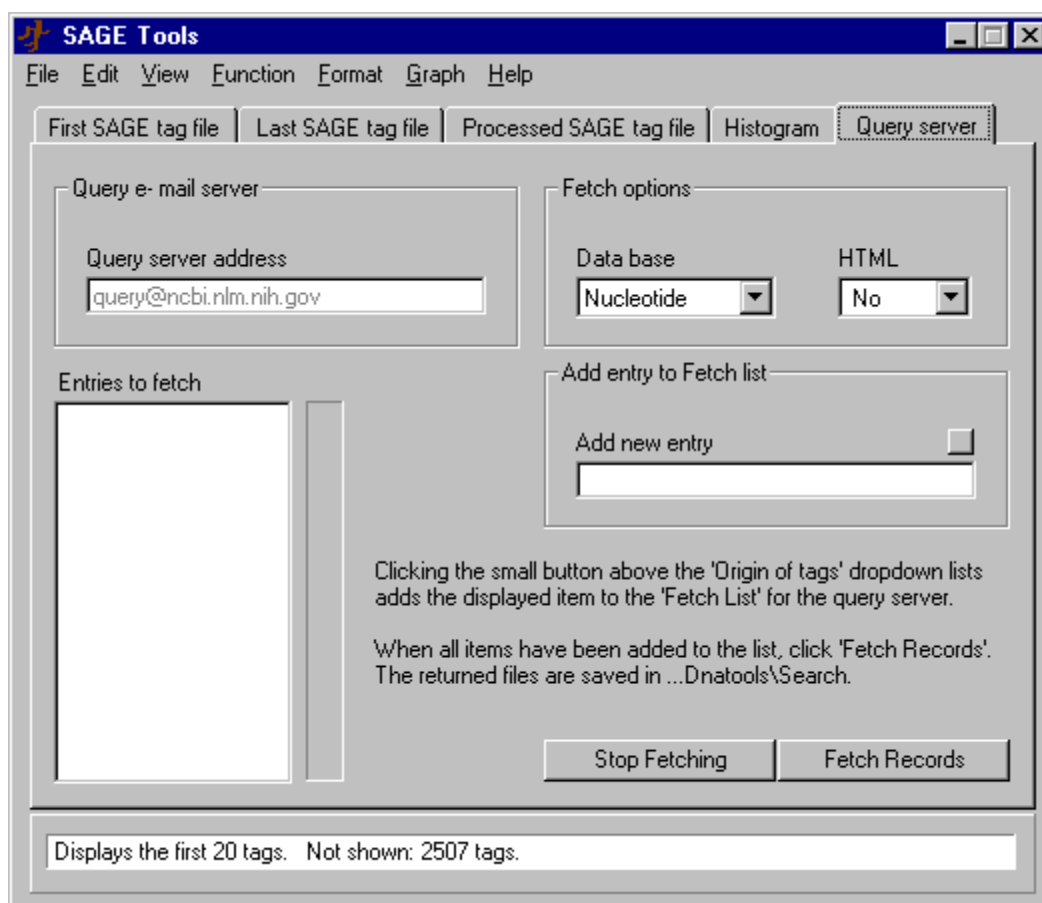
图选项:

这允许用户选择包含多少 SAGE 标记到图展示中或者什么时候打印分布柱状图。柱状图可被保存为*. bmp 或 *.wmf 文件或拷贝到剪贴板中。



询问服务器:

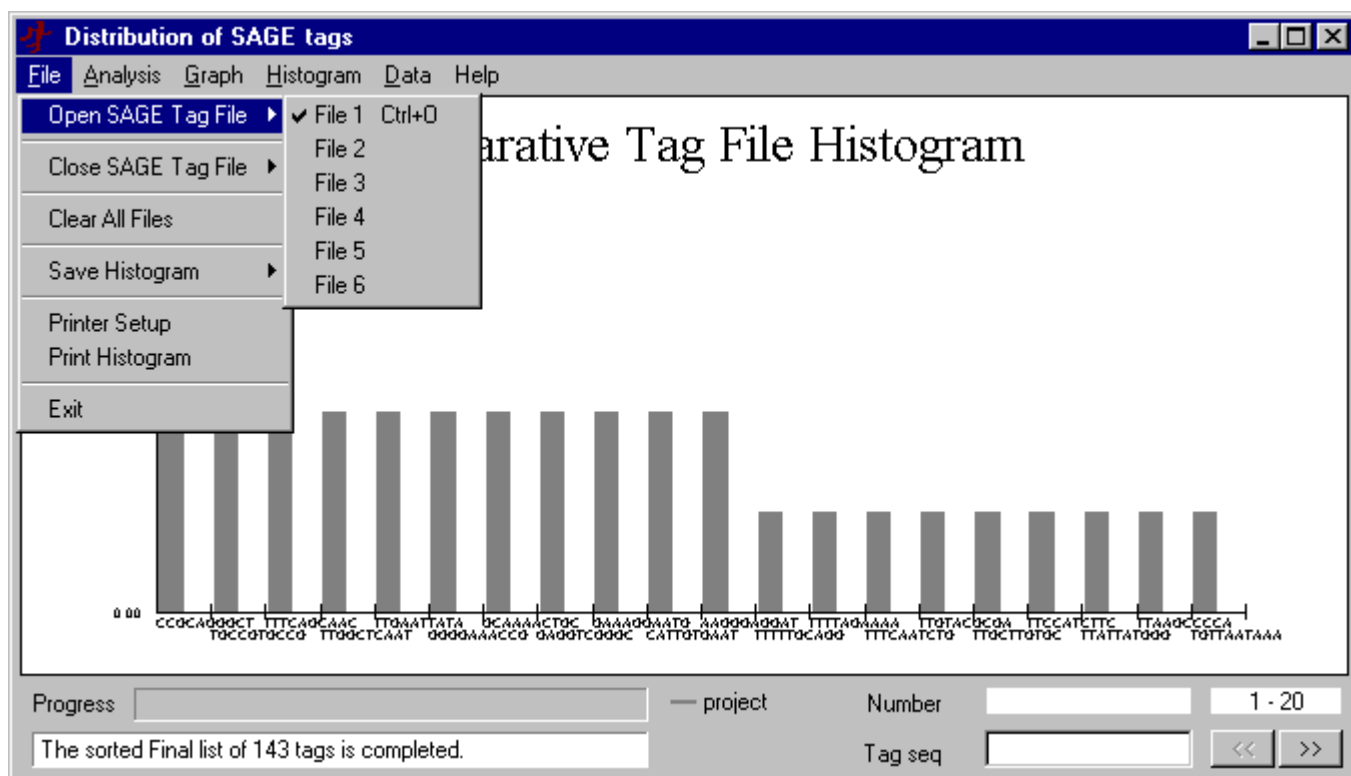
DNAtools 询问服务器利用 NCBI (National Center for Biotechnology Information) Entrez Query Server (query@ncbi.nlm.nih.gov) 提供的服务。发送一封包含数据库名字和索取号的 E-mail 到询问服务器返回一封含初始提交的电子邮件。可以从核苷酸, 蛋白质或 Medline 数据库中以此种方法提取数据。



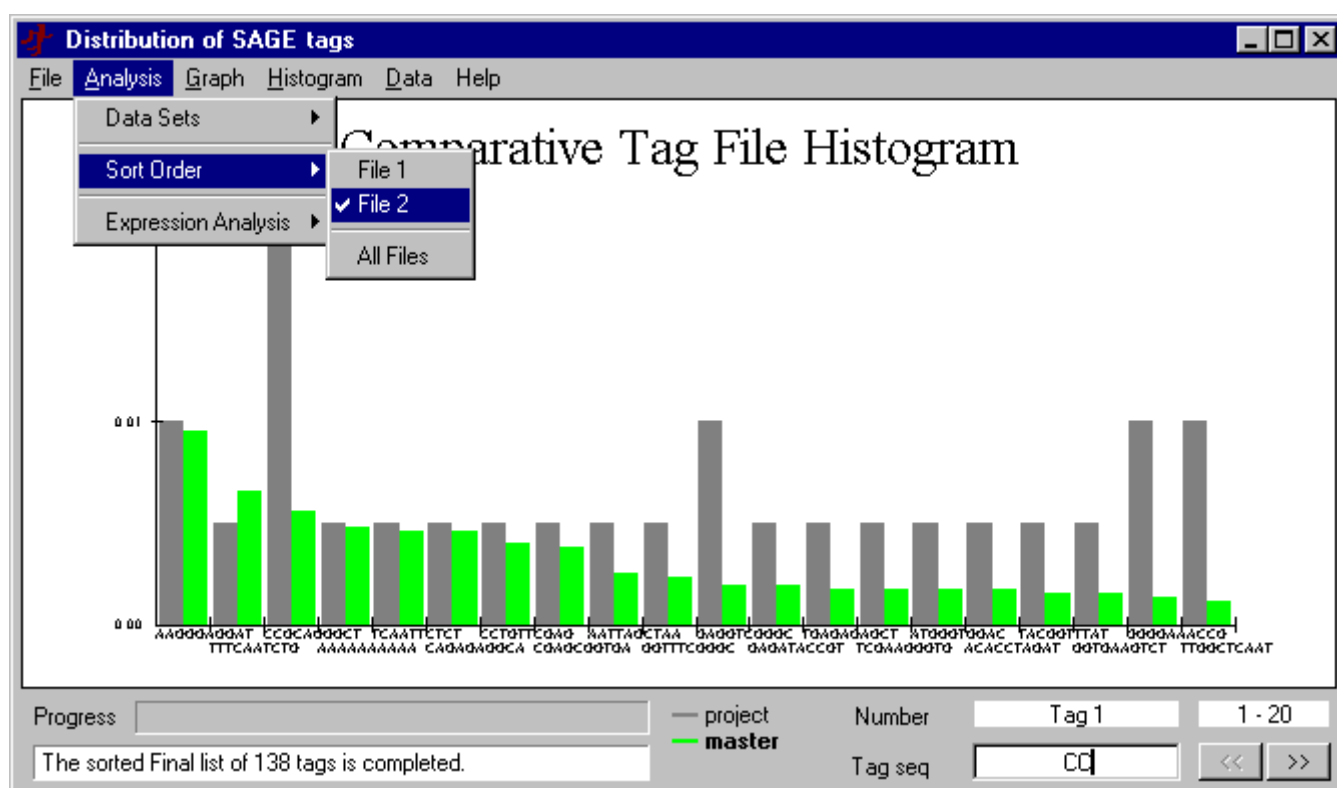
使用此功能，很容易重新寻回相应于两个 SAGE 列表比对的初始数据库入口。点击在“Origin of tag combo boxes”上的小命令按钮。这将增加选定的索取号入询问服务器表中的索取列表。这张列表也包含一个文本域用于手动增加入口。当列表完成时点击“Fetch Records”。DNAtools 会自动的发送电子邮件，直到应答返回到用户信箱并保存数据库入口到主 DNAtools 目录下的结果子目录中。

3. SAGE 轮廓分析：

此功能允许用户比较最多六个 SAGE 标记文件（用同一锚序列产生的）。比较结果被显示为柱状图或作为数据文本输出（以标记频率或起源表格的形式）。频率表格则被格式化，因此他们可被输入进一个扩展片中以进一步加工。此功能操作 DNAtools 产生的 SAGE 标记文件，扩展名*.stf (Sage Tag File)或以*.mtf format (NCBI)输入的标记文件。



第二个标记文件被装载。此张柱状图展示两个标记文件的标记频率。



此功能允许用户分析包含在一个独特标记文件中的标记是比那些首次装载的文件中的标记多还是少。换句话说，看看一个指定的基因（SAGE 标记代表的）是上调还是下调表达（在此阶段—文件 1 中的 SAGE 标记被收集）。

包含此功能的选项看上去很复杂，但是花一些耐心和实践，用户将会意识到此功能在分析大量的 SAGE 程序产生的数据时将非常有效。

为了更好的利用注释选项，用户得创建/下载 SAGEmap (*.smf) 并使用这些数据以鉴定相应于 SAGE 标记的基因。

分析：

创建最终的标记列表：

按照以下描述的进行比较：文件 1 被装载并被用于生成一个最终的加工后文件（它包含文件 1 的全部内容）。当下一个文件被装载时，主文件被升级以只包含 SAGE 标记（它同样也被包含在文件 2 中）。接下来的标记文件以相同的方式与最终的标记文件进行比较。这暗示：最终的标记文件，在装载最后的文件之后，包含 SAGE 标记（与所有装载的标记文件相同）。然后，主标记文件按照每个标记的频率总和进行分类。

数据集：

包含所有的数据集—暗示：那些不存在于所有的装载标记文件中的标记同样也被包含在最终的文件中并且被展示为柱状图，同时被包含在数据列表中。

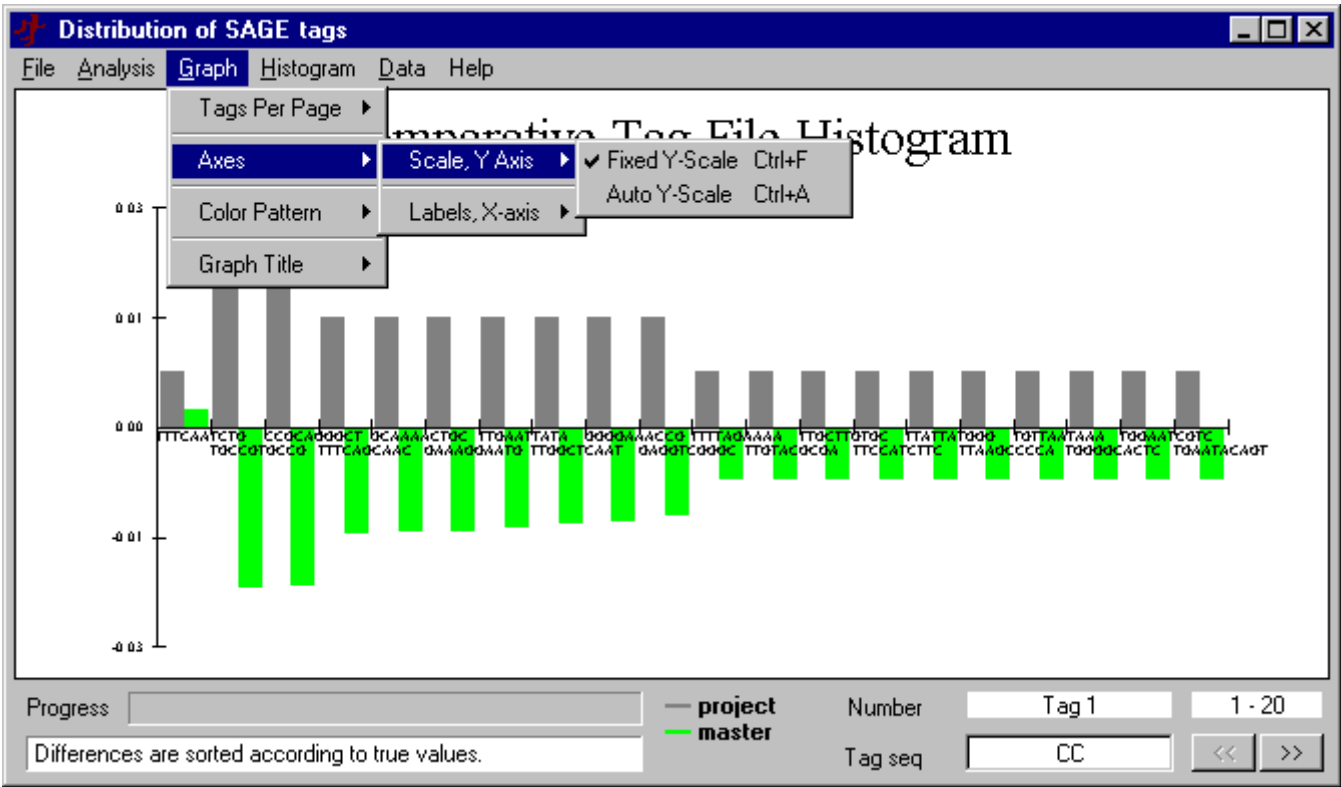
拒绝不完整数据集—若只有那些存在于所有的装载标记文件中的标记被包含在最终的文件中并且被展示为柱状图，同时被包含在数据列表中。

分类次序：

用户可以选择哪个装载的标记文件可被用于分类最终的数据。选定的分类选项应用于柱状图和数据列表。

表达分析：

“表达分析”从首次装载的标记文件中提取选定的文件(2 - 6)。阳性值暗示下调，负值暗示上调。这些差别要么依据其数值、绝对频率被分类和展示，要么依据其真值、次序频率被分类和展示。



图谱：

每页的标记：

设置展示在每个柱状图页中的数据集的数值。“Page Up, Page Down, Home 和 End”控制页之间的切换。当展示表达分析时，第一次按下 Home 或 End 将移动到阳性值和负值之间的边界。第二次，the it jumps all the way home or to the end? 当前展示的数据范围被显示在右边的域中。

轴：

比例—在此菜单中，用户可以选择柱状图是以固定的 Y 轴展示，还是调整每页柱状图的 Y 轴。当用户希望扩大标记频率之间的微小差异时，后者选项是很方便的。

标签—允许用户打开或关闭 X 轴上的标签。

颜色模式：

允许用户选择一些颜色模式用于柱状图展示。在“Visual Basic graph engine”中的选项不是主导的，但可以获得不同颜色用于所有的六个数据集。

图谱标题：

字体：

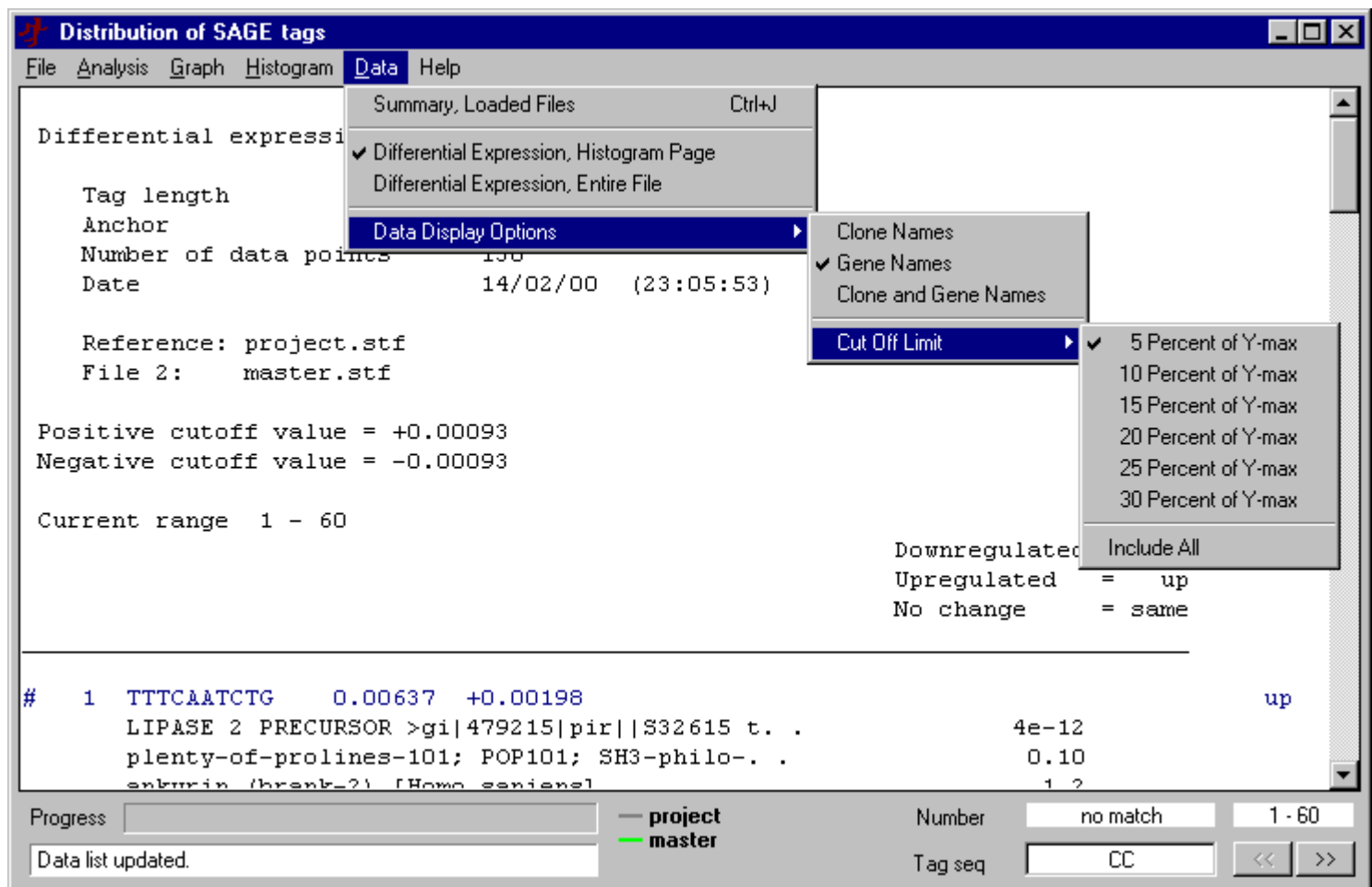
在有限的范围内改变字体大小。“graph engine”确定选定的大小匹配展示的柱状图—有必要时调节它。

标题文本：

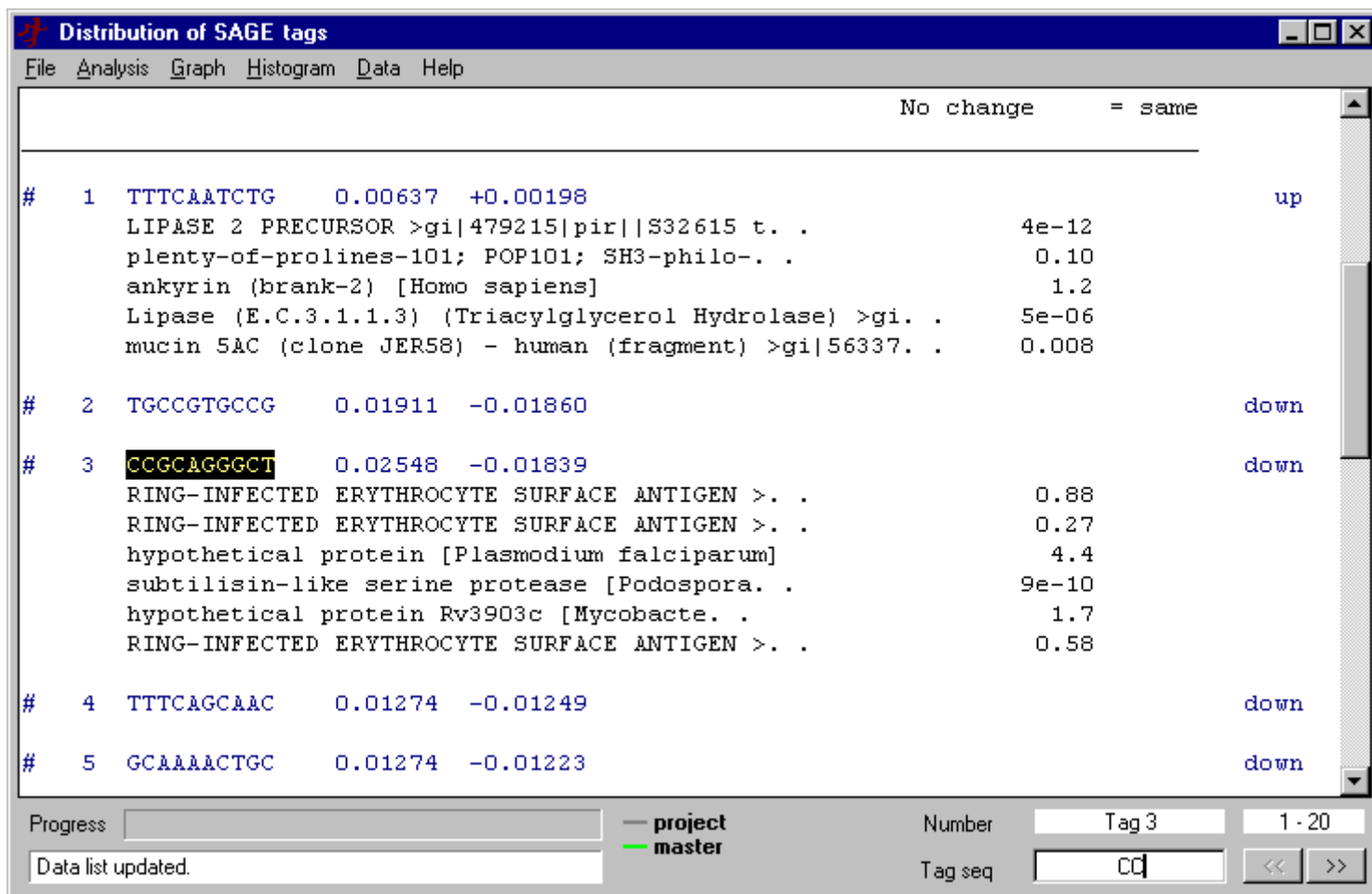
输入一个标题用于柱状图。

数据：

用于两个标记描述的比较的数据单。用于数据展示选项的菜单是展开的。



数据单的例子。DNAtools 的最新版本被稍微改变了，现在可以展示这两个标记描述中的每个标记描述的标记频率的统计分析结果。用一个来自完全注释的 EST 数据库所生成的可靠图文件注释 SAGE 标记。文件格式与从 NCBI 下载的标记和图文件是相容的。



总结装载的文件:

列出一张总结列表——总结所有装载的 SAGE 标记文件。

频率, 柱状图/完整的文件:

建造一个压缩的标记文件, 可用于当前展示的范围或用于整个数据集。数据可被装载入控制单中用于进一步加工。

起源, 柱状图/完整文件:

若序列标题数据是可获得的, 建造一个注释的标记列表。此张列表要么包含当前展示的范围, 要么整个数据集。

差异表达, 柱状图/完整文件:

建造一个包含（标记文件 1 频率）和（文件 1 和选定的标记文件之间的差异）的注释列表。下调的标记/基因用“（ - ）”注释，上调的用“（ + ）”注释。使用数据注释菜单中的终止选项，可以减少数据集的数值。

数据注释选项：

克隆名字—使用数据列表中的克隆名字而不是序列标题行。

基因名字—从 SAGEmap 文件中寻回。

克隆和基因名字—从 SAGEmap 文件中寻回。

终止限制—当比较标记文件（表达分析）时，此选项才是激活的。依据设定的终止值截短数据列表。限制是以数字式的最短 Y 轴的百分率(5, 10, 15, 20, 25, 30)进行计算的。包含所有的，设置终止值为 0，也就是无限制。在最新的 DNAtools 版本中的终止值是基于两个描述中的标记频率的统计分析。

Distribution of SAGE tags									
File Analysis Graph Histogram Data Help									
No change = same									
#	1	TTTCAATCTG	0.00637	+0.00198					up
		LIPASE 2 PRECURSOR >gi 479215 pir S32615 t. .					4e-12		
		plenty-of-prolines-101; POP101; SH3-philos. .					0.10		
		ankyrin (brank-2) [Homo sapiens]					1.2		
		Lipase (E.C.3.1.1.3) (Triacylglycerol Hydrolase) >gi. .					5e-06		
		mucin 5AC (clone JER58) - human (fragment) >gi 56337. .					0.008		
#	2	TGCCGTGCCG	0.01911	-0.01860					down
#	3	CCGCAGGGCT	0.02548	-0.01839					down
		RING-INFECTED ERYTHROCYTE SURFACE ANTIGEN >. .					0.88		
		RING-INFECTED ERYTHROCYTE SURFACE ANTIGEN >. .					0.27		
		hypothetical protein [Plasmodium falciparum]					4.4		
		subtilisin-like serine protease [Podospora. .					9e-10		
		hypothetical protein Rv3903c [Mycobacte. .					1.7		
		RING-INFECTED ERYTHROCYTE SURFACE ANTIGEN >. .					0.58		
#	4	TTTCAGCAAC	0.01274	-0.01249					down
#	5	GCAAAACTGC	0.01274	-0.01223					down

Progress
Data list updated.
project master
Number Tag 3 1 - 20
Tag seq Cc

搜索选项:

搜索标记数据—分析表的右下部分包含用于输入搜索行的文本域。此行可以是一个序列或是一个文本行。搜索安以下工作:

标记序列:

开始键入标记序列 (非前诉的锚序列)。该程序重新寻回第一个标记 (与文本域中的碱基匹配) 同时高亮显示数据单中的整个标记序列。当输入更多的碱基, 匹配会更好一直到找到标记。

自由文本搜索:

输入 “!” 以提示搜索是自由文本。然后开始输入。程序重新寻回与文本域中碱基匹配的单词同时高亮显示它。当输入更多的碱基, 匹配会更好一直到找到标记。按下 F3 寻找数据表中下一行所发生的事。

保存:

柱状图可被保存为 bitmap (*.bmp) 文件或 Windows metafile (*.wmf)。数据可被保存为 DNAtools 列表文件 (*.lst)。数据文件不能被 DNAtools 装载, 但可以在下一个编辑器或扩展单中打开。

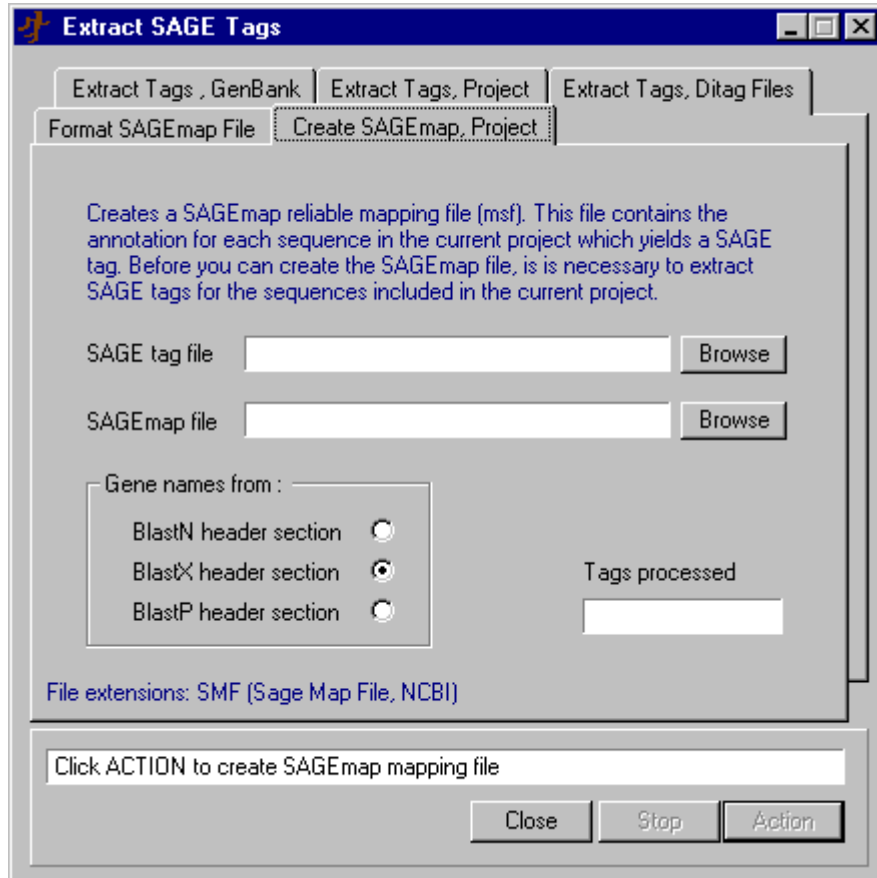
打印:

4. SAGE 可靠图谱文件:

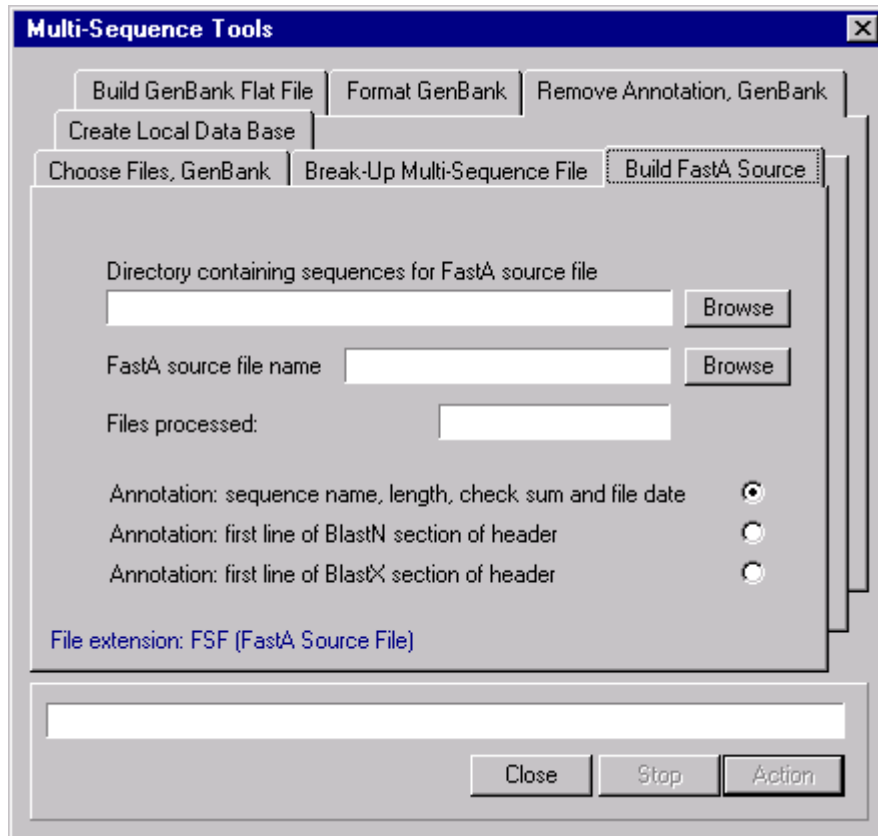
为了利用注释选项, 用户需要生成/下载一个 SAGEmap (*.smf) 并且使用这个文件中的数据以鉴定相应于 SAGE 标记的基因。

基本上, 一个位图文件是一标号定义的 ASCII 文件 (包含标记序列), 独特的基因标记子/克隆名字和注释行/基因名。

位图文件可以从 NCBI 下载, 人类的、小鼠的、大鼠的和 *S. cerevisiae* 都可以。如果用户对另一个有机体进行研究, 可以在注释的 EST 库基础上构建一个位图文件或者从一个 FastA 多序列文件中构建一个位图文件。DNAtools 包含多项功能可以用于生成来自两种数据类型的 SAGEmap。



如果不能获得 FastA 多序列文件，可以用多序列功能生成一个。



Example of a SAGemap reliable mapping file

```

AAAAAAAAA C00196-R heat shock protein 70 [Trichophyton rubrum] 2e-16
AAAAAAAAA C00224-F protein associated with DNA helicase/prim. . 6.0
AAAAAAAAA C00280-R hypothetical protein Rv2052c [Mycobacte. . 0.37
AAAAAAAAA C00822-M HYPOTHETICAL 24.1 KD PROTEIN C17A5.08 IN CH. .
9e-19

AAAAAAAAA C01407-R No description list for sequence C01407-R.
AAAAAAAAA C0A12-1R mucin, tracheobronchial - dog
>gi|402558|emb|CAA4891. . 8.5

AAAAAAAAA D00131-F No description list for sequence D00131-F.
AAAAAAAAA D00369-F 64aa long hypothetical protein [Aerop. . 0.008
AAAAAAAAA D00428-R No description list for sequence D00428-R.
AAAAAAAAA D00470-M No description list for sequence D00470-M.
AAAAAAAAA D00581-F HEAT SHOCK PROTEIN HSP1 (65 KD IGE-BINDING . .

```

6e-44

AAAAAAAAAA D00599-M No description list for sequence D00599-M.

AAAAAAAAAA D00620-F TYPE II DNA MODIFICATION ENZYME (METHYLTRA. .

0.36

AAAAAAAAAA D00762-M HYPOTHETICAL 37.2 KD PROTEIN IN ALG9-RAP1 I. .

6e-04

AAAAAAAAAA D00818-F No description list for sequence D00818-F.

AAAAAAAAAA D00837-F PUTATIVE GLUCOSYLTRANSFERASE C08H9.3 >gi|38. .

6.3

AAAAAAAAAA D00940-M endonuclease [Magnaporthe grisea] 5e-53

AAAAAAAAAA D01107-M A2-5a orf23; hypothetical protein [Ba. . 1.7

AAAAAAAAAA D01268-F GTP-binding protein ypt5 - fission yeast

(Schizosacc. .2e-12

AAAAAAAAAA D01294-M glycoprotein [Vesicular stomatitis virus] 9.5

AAAAATCTTG D00950-M LONG-CHAIN-FATTY-ACID--COA LIGASE 3 (LONG-C. .

7e-10

5. 用 SAGE 轮廓寻找方案:

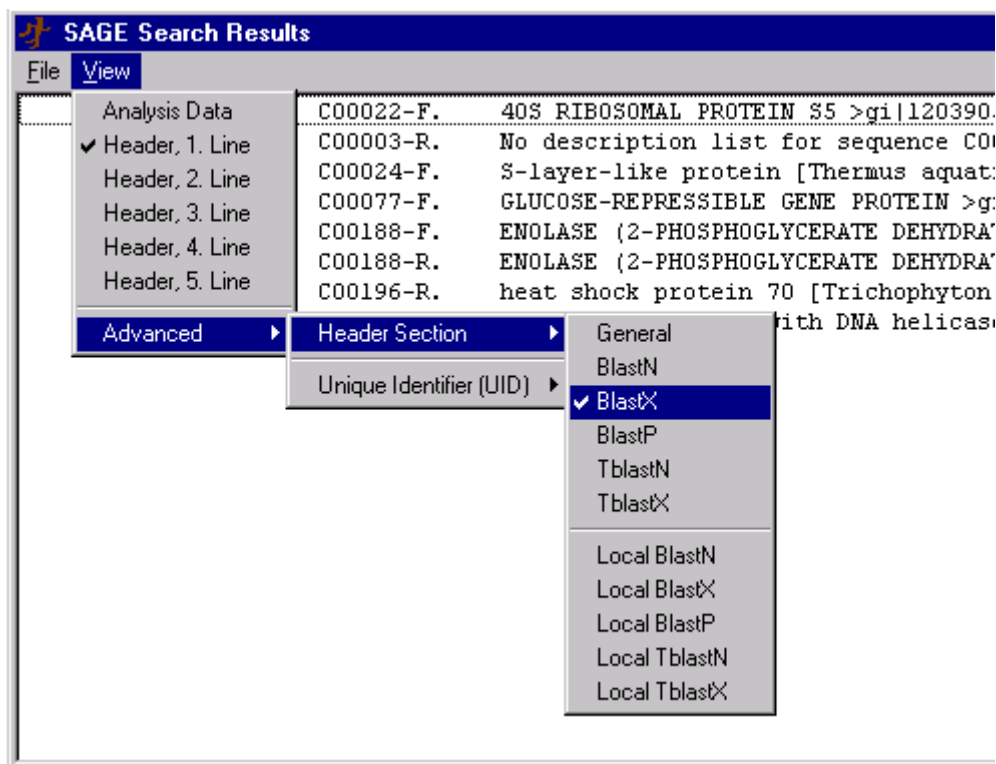
SAGE 分析功能被用于搜索当前装载于方案中的所有的序列文件 (SAGE 标记列表, 例如 9—12 碱基对短的序列) (Serial Analysis of Gene Expression, Victor E. Velculescu, Lin Zhang, Bert Vogelstein, Kenneth W. Kinzler, Science 270, pp484-487 (20 October 1995))。

在启动搜索前, SAGE 标记数据文件*. stf 或*. ptf 必须被装载“*File/Open Sage Data File*”。在打开数据文件前, 输入 SAGE 标记的长度。所有的标记序列被核实 (只有 ACGT 被允许) 并与指定的长度比较。在装载 SAGE 标记数据文件的过程中, 那些不同长度的标记和或包含不符合规定的字符的标记将被拒绝掉。

若用户希望在搜索中包含一部分或所有的锚序列, 输入希望包含在锚序列域中的那些碱基。

若用户希望搜索两条链、沃森或克里克链, 在 “strand” 中选择。

搜索结果被展示在一个独立的表中并包含序列名, 标签数值, ID 和序列。锚序列与标记序列被 “/” 分开。在结果表中的查看菜单中, 用户可以指定一个标题行 (line 1 to 5) 同时还可以将结果作为一个指定的标题行列表来查看。



双击在结果列表中的一行将重新寻回相应序列的标题。关闭标题表返回到 SAGE 分析结果列表。

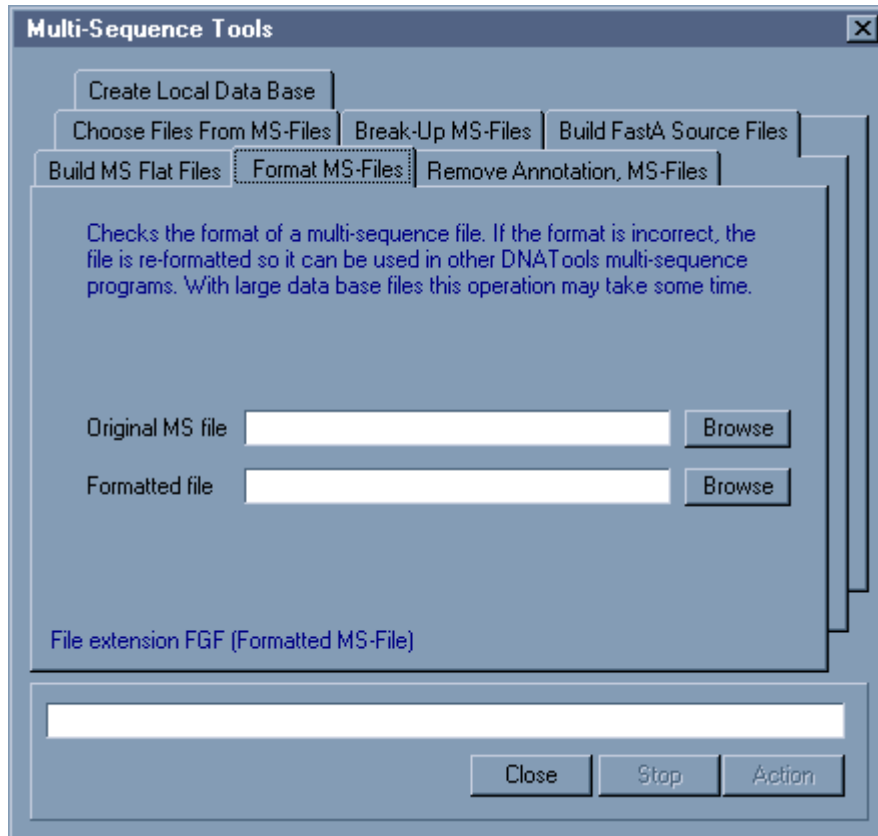
发给使用者的信息被展示在 info 行（信息行，在分析表的底端）。

6. 多序列工具：

此项功能允许用户修改或生成多序列文件。FlatFile 或 FastA 格式的数据库文件包含文本文件，而这些文本文件又含纯 ASCII 格式的多数据库且此数据库文件可从几个 WEB 站点如 NCBI 处下载到。（<ftp://ncbi.nlm.nih.gov/genbank/>）

格式：

此项功能审查多序列文件的 EOL(end of line)码以确认 LF 是否单独使用。如果的确如此，所有的 LF 码被 CRLF 码代替且格式化后的文件以相同的名字被保存，但扩展名改为 *.fgf(Formatted Genbank File)。万一正确的 CR 码在被使用，该文件的一个拷贝以扩展名为 *.fgf 被保存。



移除注释：

此项功能移除来自 flatfile 格式的 GenBank 文件中序列的注释，但保留 DESCRIPTION 和 ACCESSION 行不变。初始的分离子和//记录分离物在修改的文件中仍然被保存下来。

如果选择审查 “Exclude sequences > 15,000 bases”，那么长于 15000 碱基的序列被从修剪文件中排除掉。在大多数情况下，如此长的序列包含多基因序列，从此序列中提取的 SAGE 标记是无意义 的。

输入文件必须是扩展名为 *.fgf 或 *.ngf，如用户生成文件或一个 GenBank 文件且这些文件的 EOL 码已经被审查并且若不正确则被替换。修改的文件以相同于输入文件的文件名被保存，但扩展名改为 *.tgf。 (Trimmed GenBank File)。

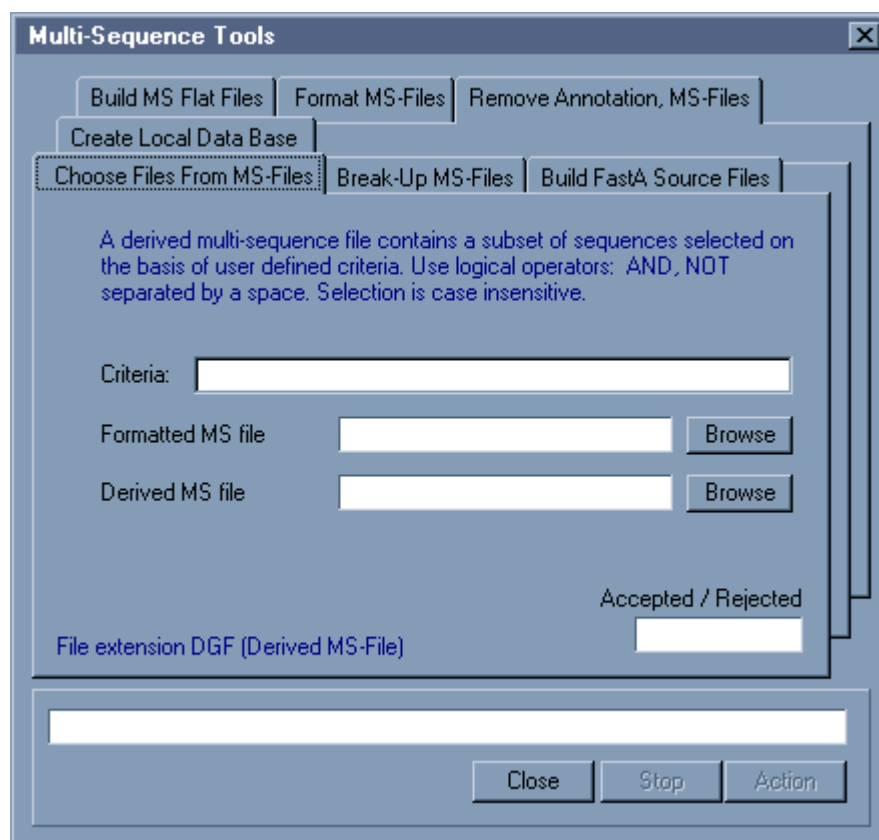
衍生 GenBank：

此项功能允许用户从格式化的 GenBank 文件中提取一亚组序列。如从完整的 EST 数据库平台文件中提取一个特殊生物的 EST 序列。GenBank 文件必须是标准的格式即：每个记录的序

列部分被 ORIGIN 和 // 所划分。注释部分必须包含 DEFINITION , ACCESSION, ORGANISM 和 REFERENCE 部分。对于 FastA 格式的多序列文件或修改过的 GenBank 文件不起作用。

标准：用于为衍生的 GenBank 文件选择文件的准则—使用逻辑算法 AND 和 NOT 输入，例子如下：

- AND Homo (retains all files containing the word Homo/HOMO/hoMo in the annotation)
- NOT Yeast NOT Saccharomy NOT cerevisiae (selects all sequences except yeast sequences)
- AND Plant (retains all plant sequences)
- AND Fungi (retains fungal sequences)



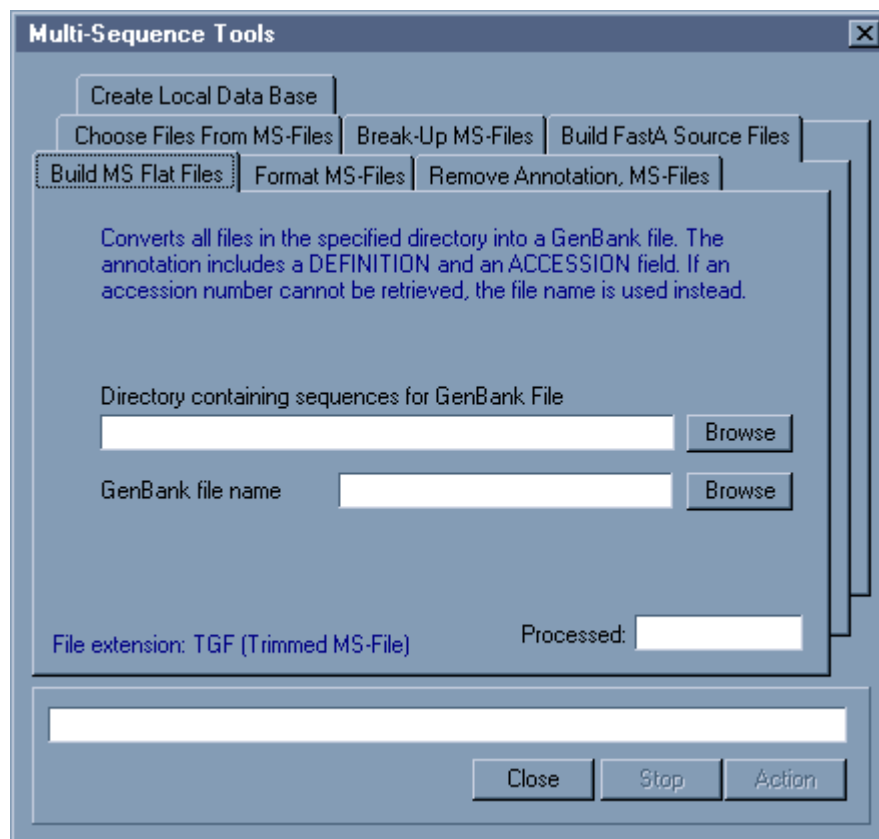
注意：在标准中，最大值只允许 5 个 AND 和 5 个 NOT。若每个类别中超过了 5 个，在选择时这些过多的词被忽略掉。逻辑算子和关键词必须用空格隔开。搜索是案例敏感的。

输入文件必须是扩展名为*.fgf 或 *.ngf，如用户生成文件或一个 GenBank 文件且这些文件的 EOL 码已经被审查并且若不正确则被替换。修改的文件以相同于输入文件的文件名被保存，但扩展名改为*.tgf。(Trimmed GenBank File)。

明显的，此项功能只与 GenBank 文件相关，此时注释没有被修改。

生成新的：

用于依据自己的序列生成一个新的 flatfile 格式的 GenBank 文件。在尝试创造文件之前，确信所有包含于文件中序列定位于相同的目录。新的 GenBank 文件的记录格式与那些修改过的 GenBank 文件相同而且新文件的扩展名为*.ngf (New GenBank File)。



唯一保留的序列鉴定信息是文件名，在 DESCRIPTION 后被包括在内。所有包含于标题中的其他信息则被丢失。

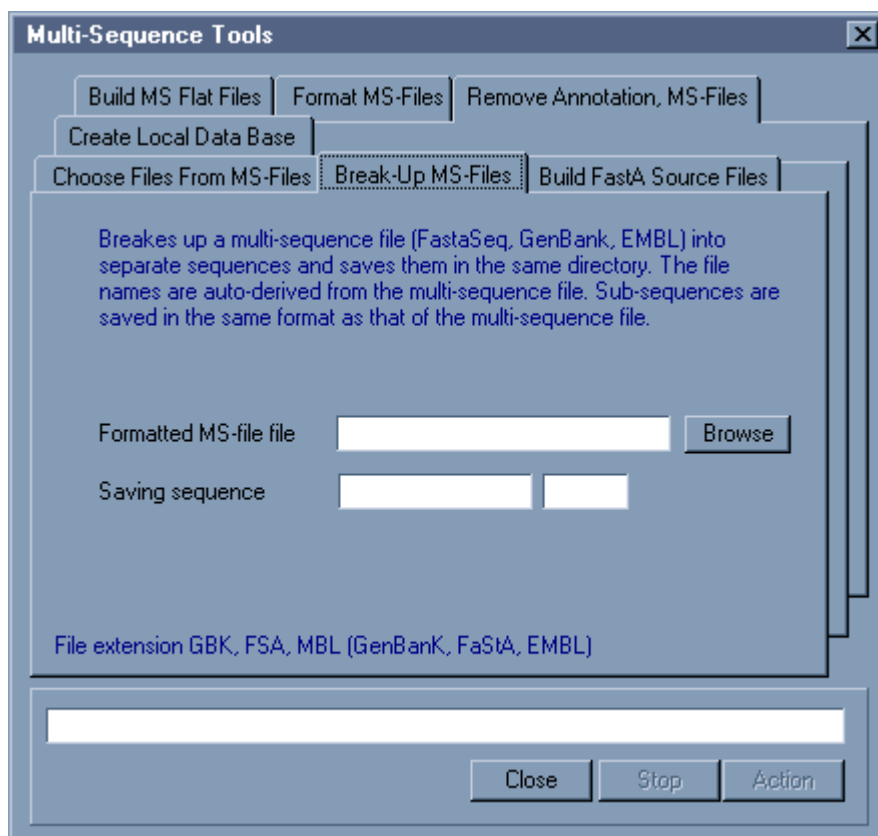
DNAtools 接受 GenBank, DNATools 和 FastA 格式的文件。在某些场合, 当整个数据库文件出现在文件标题中时, 大于一个标题/序列分割子将会出现在文件中。这将导致标题/注释和序列不正确的分离。

解决此问题的方法是装载文件入 DNAtools 中并且再一次保存他们: 在文件被保存前, DNAtools 审查每个文件的标题, 看看是否有不合规定的分离子。如果有多个分离子出现在标题中, 他们将会被转化为分离子不认识的参数。

此项功能可以被用于生成一个多序列文件以用于 SAGE 标签提取。换句话说, 用户可以装载大量的文件入方案中, 相反的, 首先将其合并成一个多序列文件然后使用 SAGE 提取功能以生成 SAGE 标签文件。

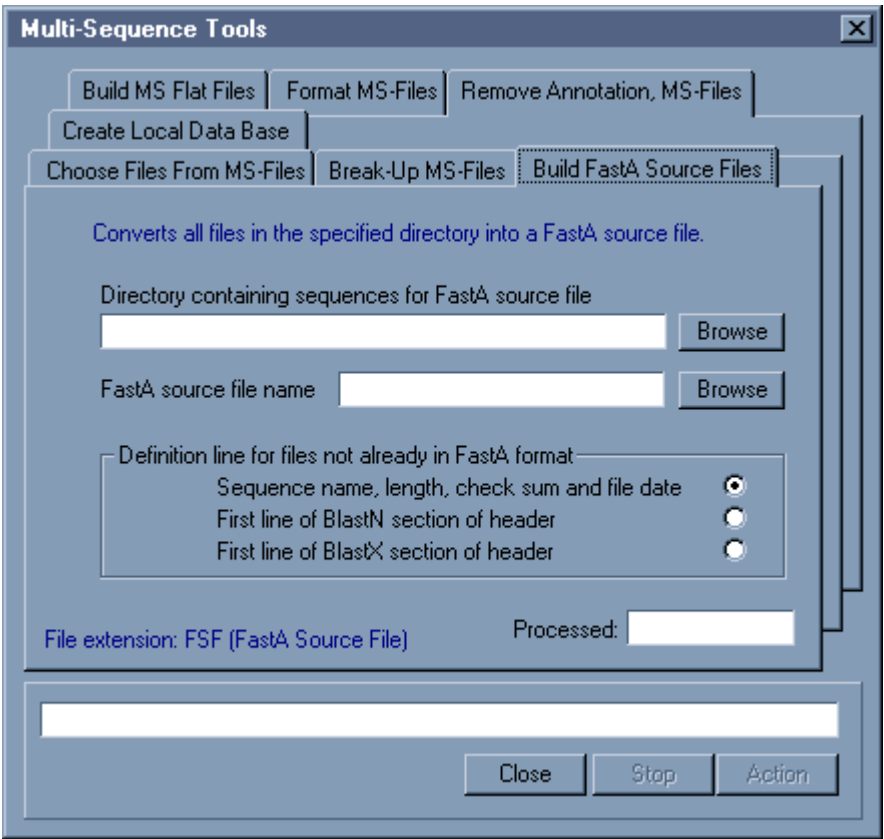
打碎多序列文件:

此项功能打碎一个 GenBank 或者 FastA 格式的多序列文件同时在相同的目录下保存每个子文件。子序列的文件名由这些子序列的索取号组成, 扩展名为*.gbk 的是用于 GenBank 文件而扩展名为*.fsa 是用于 FastA 文件。



生成 FastA 源文件：

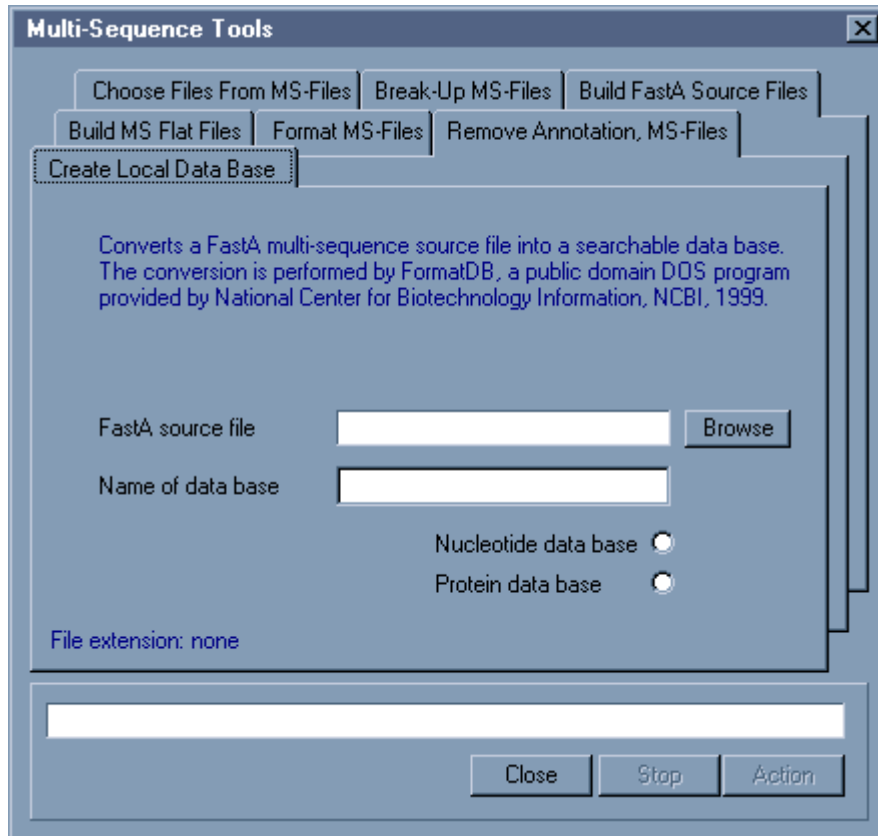
为了生成一个 FastA 源文件，有必要收集所有的特定序列（即用户期望这些序列包含在同一目录下的源文件中）。如果 DNAtools 已经将这些序列进行注释，用户有必要提示用户希望哪个标题部分 (blastN or blastX) 用于注释这个源文件。



生成本地数据库 (with formatdb, NCBI)：

此项功能运用 NCBI 的 DOS 程序 FormatDB 以生成可搜索的数据库。

查看 “*Local Blast Search*” 寻找 formatDB DOS 程序的安装信息。为了生成一个数据库，按照一下步骤：



如何生成一个本地数据库：

拷贝所有的用户希望将其包含在本地数据库中的序列到一个空的目录下；

点击 “Utilities/Multi-Sequence Functions/Build FastA Source” 从目录中的文件以生成一个 FastA 源文件；

如果序列文件是 DNAtools 格式，选择注释源文件-否则注释将自动的从源文件中被提取出来；

点击 “Utilities/Multi-Sequence Functions/Create Local Data Base” 以建造本地数据库；

记住：如果正确的数据库类型被选择了，审查核实一下；

完成的本地数据库被创建在 Windows / Winnt 目录下的 DT5_TEMP 子目录中，同时如果希望搜索它时必须将其保留在那里；

万一希望从含 DNA 序列的方案中生成一个蛋白质数据库，在建造 FastA 源文件和最终的数据库之前，使用“Utilities/Create protein files”翻译该核苷酸序列。

注释：该数据库中序列将只包含一行注释。如果输入序列是 GenBank, FastA 或 GCG 格式，注释将被自动的从初始文件中寻回。对于 GenBank 文件，使用 DESCRIPTION 行；对于 FastA，使用标准的单行注释；对于 GCG，.. 分割符号之前跟行。

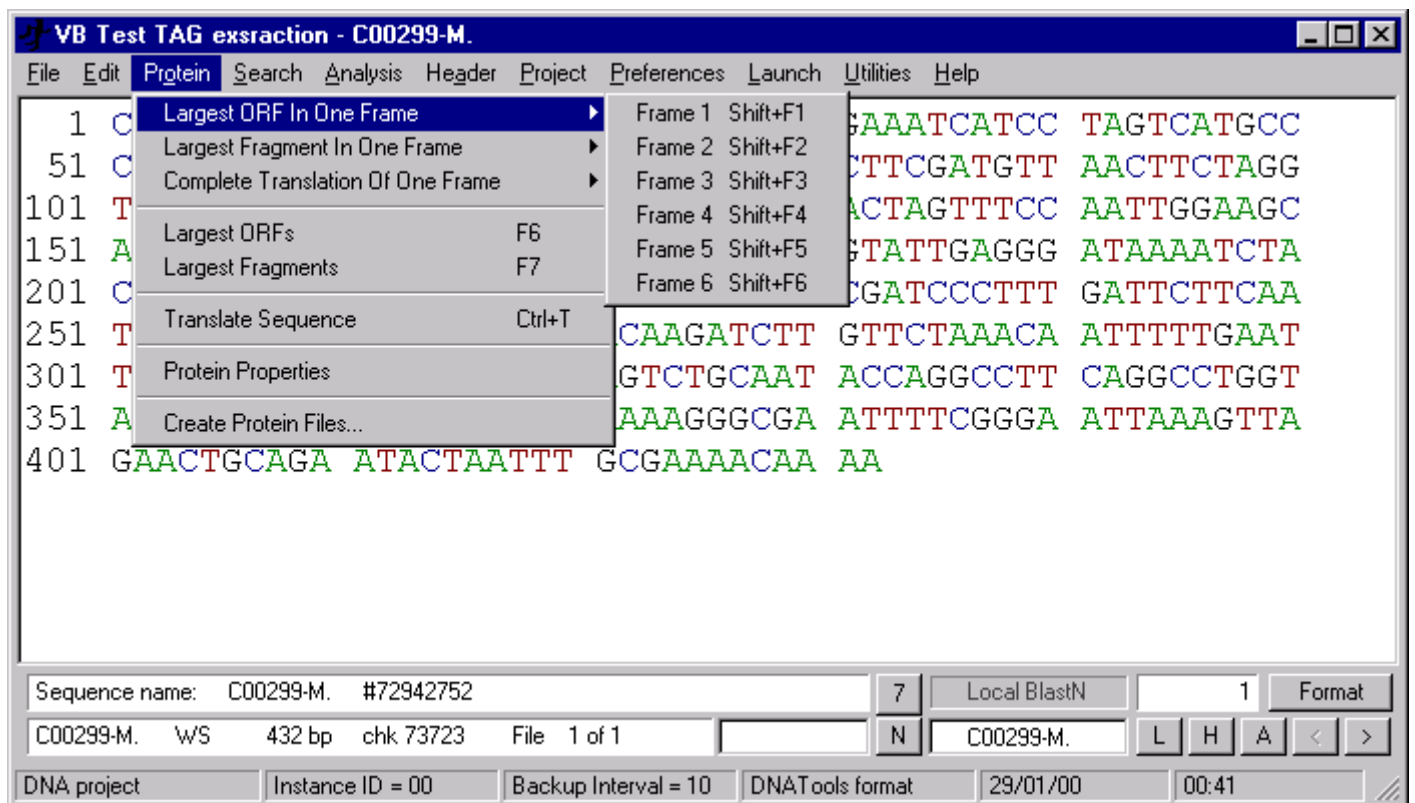
对于 DNAtools 文件，列在表上的标题部分选择哪个部分用于注释。默认的是序列名，长度，审查总和和文件日期。如果文件标题包含 blastn 和或 blastx 搜索结果，其中的一个标题部分的第一 DESCRIPTION “描述” 行可被选择用于数据库入口的注释。对于一个或多个序列标题，如果选择的标题部分 (Blastn: or Blastx:) 是缺失的，将使用默认的设置。

万一用户的序列包含一个 Blastn 和一个 Blastx 标题部分，用户可以用 blastn 注释生成一个数据库并且可以用 blastx 注释生成第二个数据库。

Chapter6: DNAtools-translate options

1. 翻译当前 DNA 序列：

可以搜索序列中的 ORFs 或片断（所有六种读码框）或在每个读码框中的最长的 ORF 或片断。搜索结果可被保存或打印。最长片断的相配之物显示于序列的状态行。

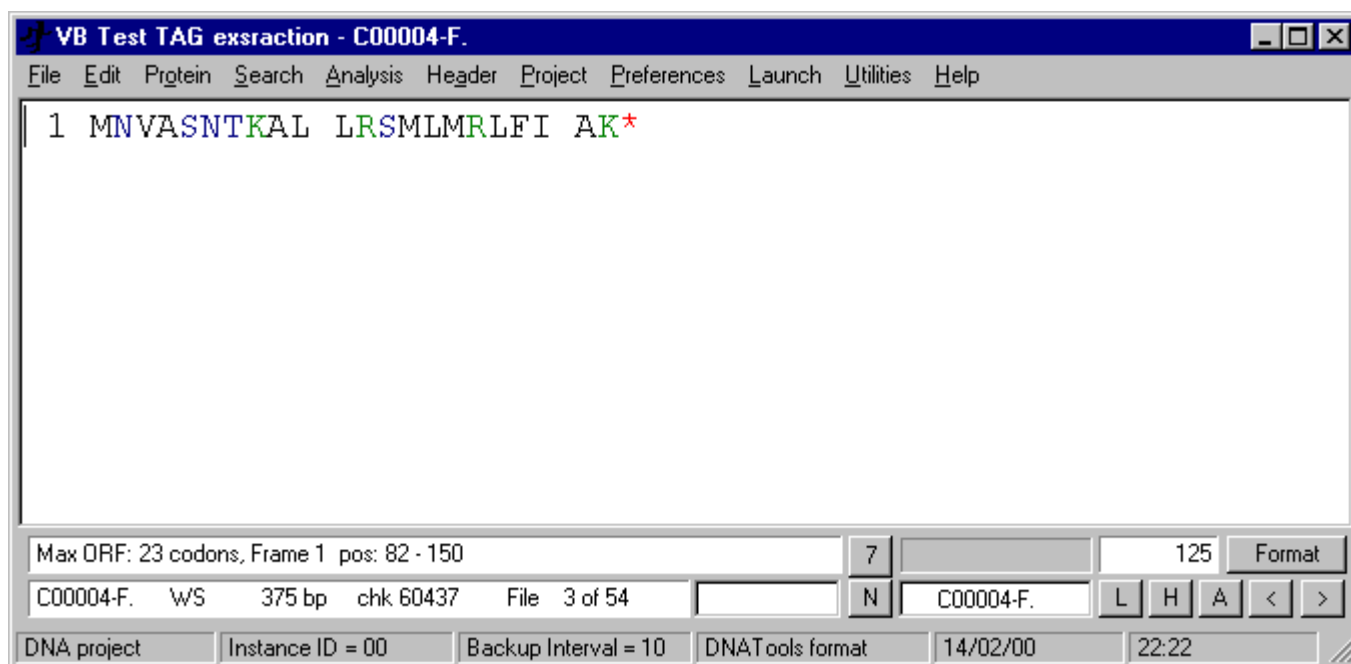


完全翻译，一个读码框：

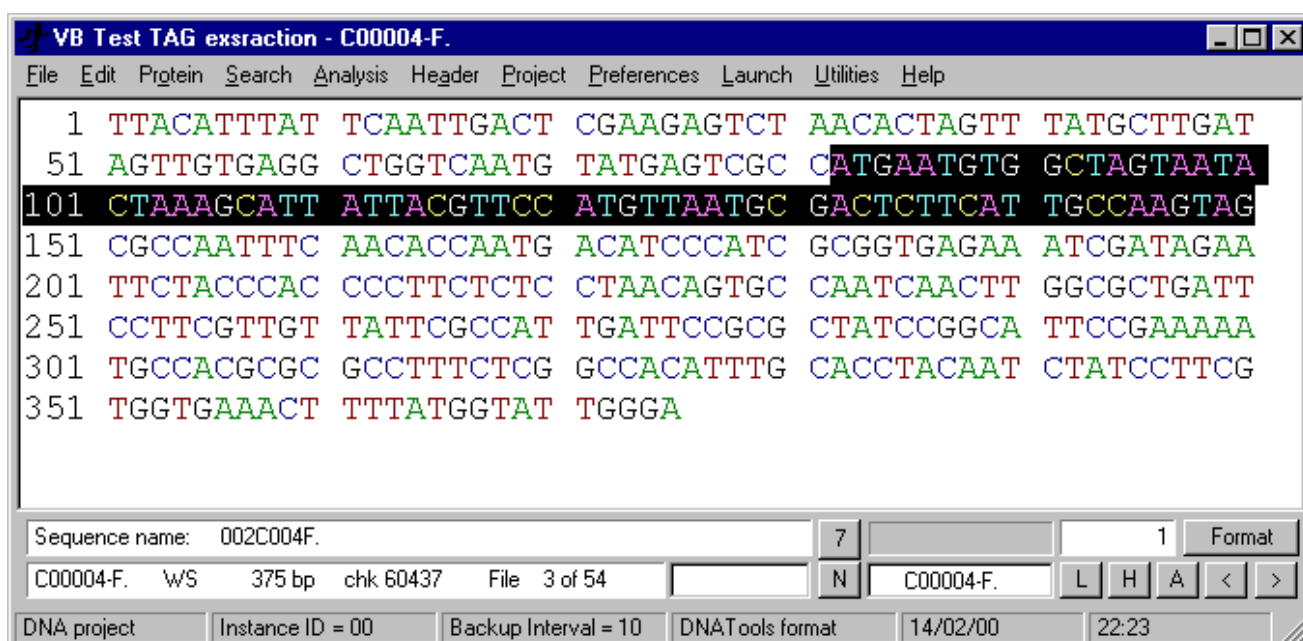
此功能“Protein/Complete Translation of One Frame”翻译完整的 DNA 序列（以指定的读码框）同时展示完整的包含终止密码子的蛋白质序列（以“*”标记）。翻译的序列可被保存以用于以后的数据库搜索使用。

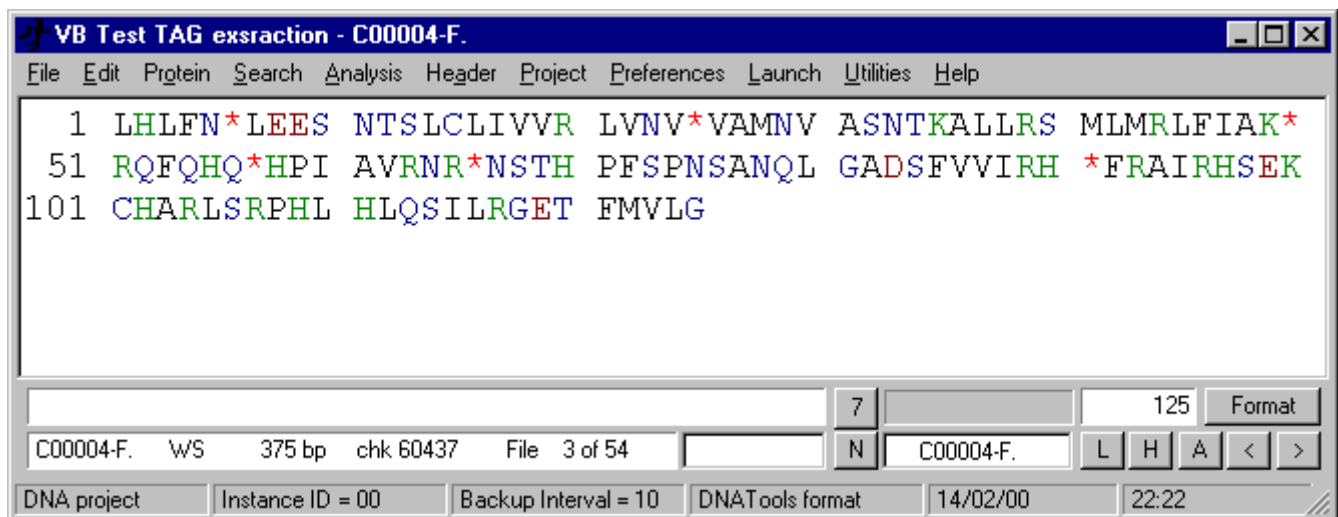
在一个读码框中寻找最长的 ORF：

此项功能“Protein/Largest ORF in One Frame”翻译 DNA 序列（以指定的读码框）同时展示最长的 ORF 和其相应的信息。所有的六个读码框的 ORFs 可以同时被展示，或者分开展示。



当翻译序列被展示时，点击格式按钮，高亮显示 DNA 序列中翻译的区域。



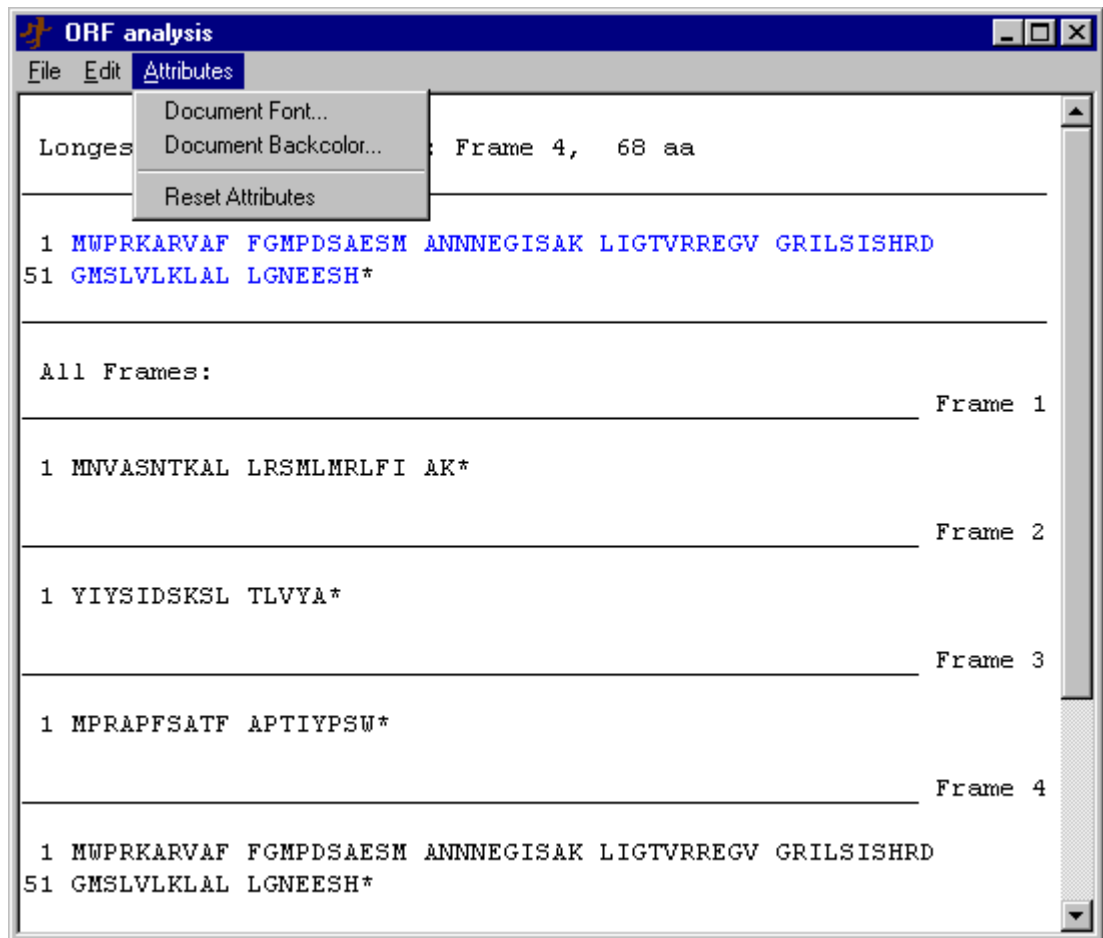


在一个读码框中寻找最长的片断:

此项功能 “Protein/Largest Fragment in One Frame” 翻译 DNA 序列（以指定的读码框）同时展示最长的未打断的蛋白质片断和其相应的信息。

在所有读码框中寻找最长 ORF:

此项功能 “Protein/Largest ORF In Six Frames” 翻译 DNA 序列同时展示最长的 ORF（每个读码框）。



在所有读码框中寻找最长片断：

此项功能“Protein/Largest Fragment In Six Frames” 翻译 DNA 序列同时展示最长的氨基酸片断（每个读码框）。

评述：略

定义：

开放阅读框（ORF）— 在一个单阅读框中的一段 DNA 序列，按照以下限定：

DNA 序列的起始或者紧跟一个终止密码子的第一个 ATG 密码子；

DNA 序列的末端或者一个终止密码子（TAA, TGA, TAG）。

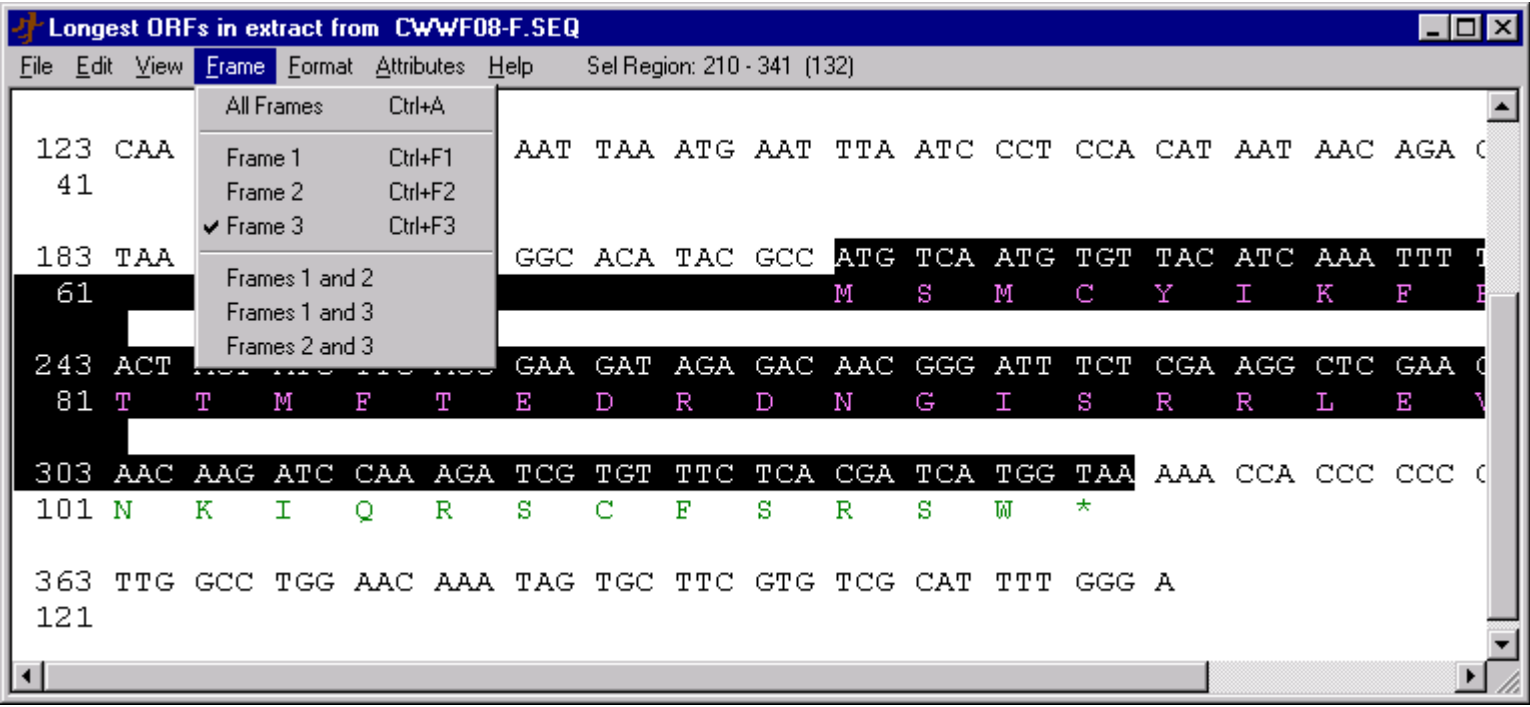
片断— 在一个单阅读框中的一段 DNA 序列，按照以下限定：

一个终止密码子或者蛋白质序列的起始；

一个终止密码子或者蛋白质序列的末端。

2. 翻译 DNA 序列图：

此 “Protein/Translate Sequence Extracts” 表展示当前 DNA 序列的翻译情况。



行数相应于提取序列区域的相配之物。终止密码子用星号注释，用 X’ 注释不确定氨基酸。

DNA 序列的格式与序列编辑表中选定的格式无关（阻断长度是 3，行长度 60bp）。

文件：

打印—打印序列和翻译（就像显示于屏幕上的一样）。

查看：

完整翻译—展示 DNA 序列完整的翻译（以选定的读码框）。

所有的开放式读码框—展示所有开放式读码框（以选定的读码框）。

最长开放式读码框—展示最长的开放式读码框（以选定的读码框）。

最长片断—展示两个终止密码子之间最长的片断（以选定的读码框）。使用此选项定位序列错误（产生读码框漂移错误）。

框架：

所有框架—在 DNA 序列之下展示三个前向读码框的氨基酸序列。

框架—翻译 DNA 序列并在 DNA 序列之下展示氨基酸序列（以选定的前向读码框）。

高亮显示一个区域：

使用锚—使用此选项，可以在氨基酸序列的基础上选择一个序列区域（如在一个开放式读码框中的一个内含子）。当表打开时，锚 1 被激活用于输入区域的底端限制。在点击选定碱基的右边之后，锚 2 被选定（或者按下 CTRL+F2 或者从主菜单中选择）。再次点击碱基的右边。点击 Pos 菜单项目或者从组菜单中选择“Anchor/Show Region”。这将展示选定的区域。当区域仍是高亮显示时，关闭表以高亮显示展示在主编辑器中的序列区域。这个区域可以被取代，拷贝或删除。

通过拖动—高亮显示序列的一部分。Mouse-Click -> Mouse-Down -> Mouse-Drag -> Mouse-Up 有相同的作用：当一个区域被高亮显示时关闭翻译表，高亮显示的部分被维持显示（当序列展示在常规序列编辑器中）。在开始拖动之前，务必抓住区域的起始部分。选定的起始位点展示在主菜单中。当区域仍然被高亮显示时，关闭表以高亮显示展示在主编辑器中的序列中的区域。区域可被拷贝、删除。

位点：

展示当前选定的锚/起始点和锚所定义的区域长度。

3. 自翻译 DNA 序列：

这个功能是为了帮助用户进行短 EST（表达序列标签）序列分析而设计的，以防通过数据库查询功能鉴定失败且正确的读码框又不知道。

这个功能以选定的读码框翻译所有当前方案中的 DNA 序列，同时保存每个蛋白质序列为独立的文件。这些蛋白质序列可以针对蛋白质基序的数据库或者其他包含蛋白质信息的数据库进行查询。

Create Protein Files

Translation options

Complete translation

Longest fragment

N-terminal fragment

C-terminal fragment

Frame options

All frames

Forward frames, 1, 2, 3

Reverse frames, 4, 5, 6

Filter options

Minimum length060

Format options

FastA file format

GCG file format

File Options

Each frame in one file

All frames in one file

Insert 5 x stop between frames

Add two-char. frame-code to

Complete file name (8 chars.)

Rightmost 6 chrs. of name

Leftmost 6 chars. of name

Middle 6 chars. of name

Overwrite without warning

C00003-R. _A11	742
C00004-F. _A11	748
C00004-R. _A11	771
C00005-F. _A11	769
C00005-R. _A11	661
C00006-F. _A11	697
C00006-R. _A11	770
C00007-F. _A11	873
C00007-R. _A11	762
C00008-F. _A11	802
C00008-R. _A11	946
C00009-F. _A11	872
C00009-R. _A11	769
C00010-F. _A11	883
C00010-R. _A11	772

The list includes 15 protein sequences longer than 60 aa

Help

Close

Destination

Create Files

翻译选项：

完全序列：完全翻译的序列包含 X 和中止子；

最大片断：最大连续的氨基酸区（没有中止子）；

N-端区：以 M 开头且以第一个下游中止子为结尾；

C-端区：从序列开始到第一个中止子的区域。

框架选项：

所有的读码框；

前向 3 个读码框；（A, B, C）

反向 3 个读码框；（D, E, F）

过滤选项：

过滤选项允许用户忽略那些短于选定最小长度的蛋白质序列。

蛋白质文件名：

可以通过在 DNA 序列文件名中加_N 来定义蛋白质文件名。N 表示读码框（1-6 或者 # 表示在同一文件中所有的读码框）。这两个参数可以以下四种形式之一进行加入编辑。

用_N 替换 DNA 序列文件的扩展名；

将_N 加到文件名最左边的六个参数；

将_N 加到文件名最右边的六个参数；

将_N 加到文件名中间的六个参数。

在后面三种情况下，蛋白质文件名缺少扩展名。当建造蛋白质文件时，选定的文件名将被核实以避免相同的文件名。

如果选定的命名方法产生相同的文件名，创建过程将被停止同时建议选择另外一种产生蛋白质文件名的方法。万一四种方法都不能产生单一的蛋白质文件名时，初始的 DNA 序列文件名需要重新命名。

文件格式：

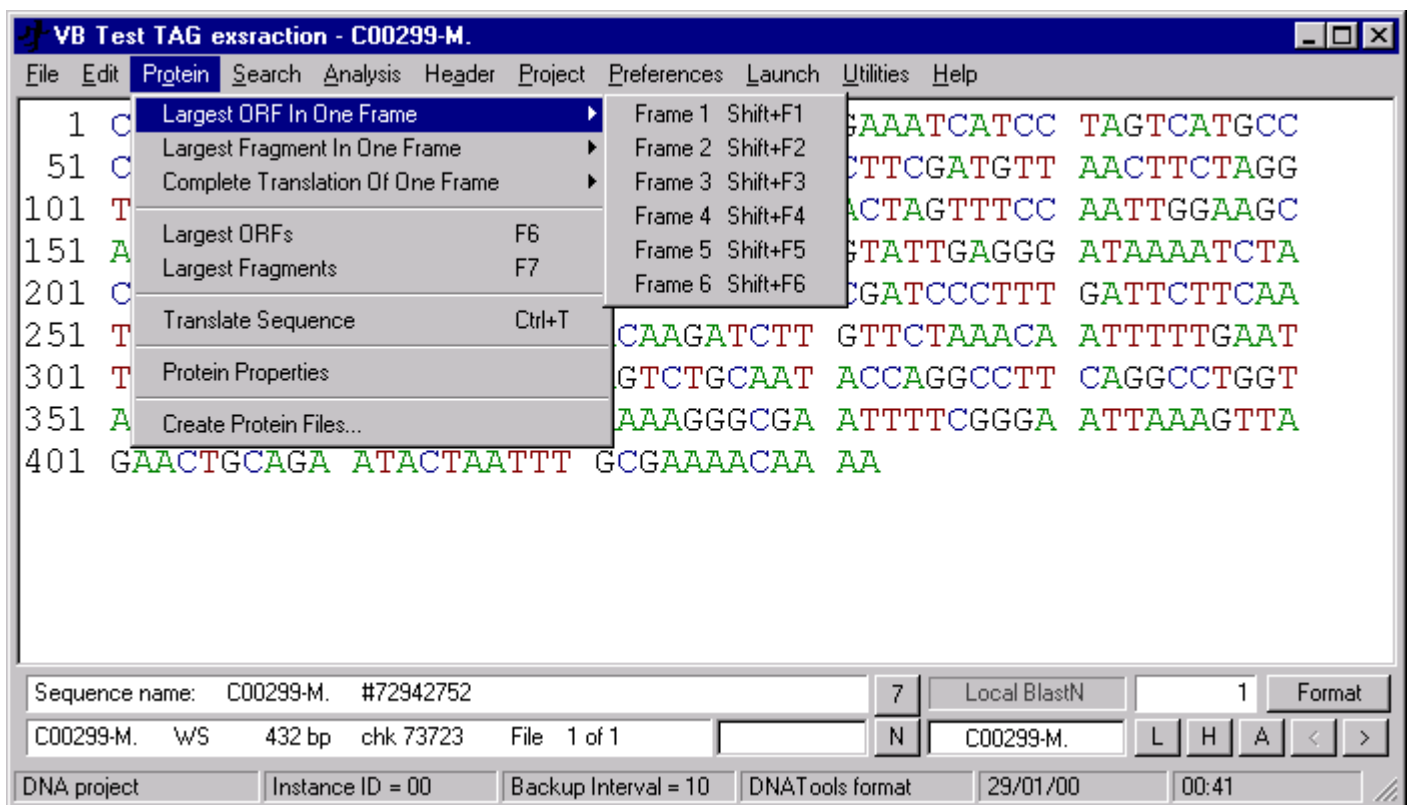
蛋白质文件可以以 Fasta 或 GCG 格式进行保存。每个蛋白质文件包含一个标题，给出序列名称，读码框和蛋白质序列的长度。蛋白质序列被打断成 50 参数的行，但没有行标记。

存储选项:

蛋白质文件既可以以独立的文件进行保存, 还可以以每个 DNA 序列保存相应的蛋白质序列文件。如果选择后者且选择审查选项框时, 在每个读码框之间将加入 5x 间隔物。

4. 翻译当前 DNA 序列:

可以搜索序列中的 ORFs 或片断(所有六种读码框)或在每个读码框中的最长的 ORF 或片断。搜索结果可被保存或打印。最长片断的相配之物显示于序列的状态行。

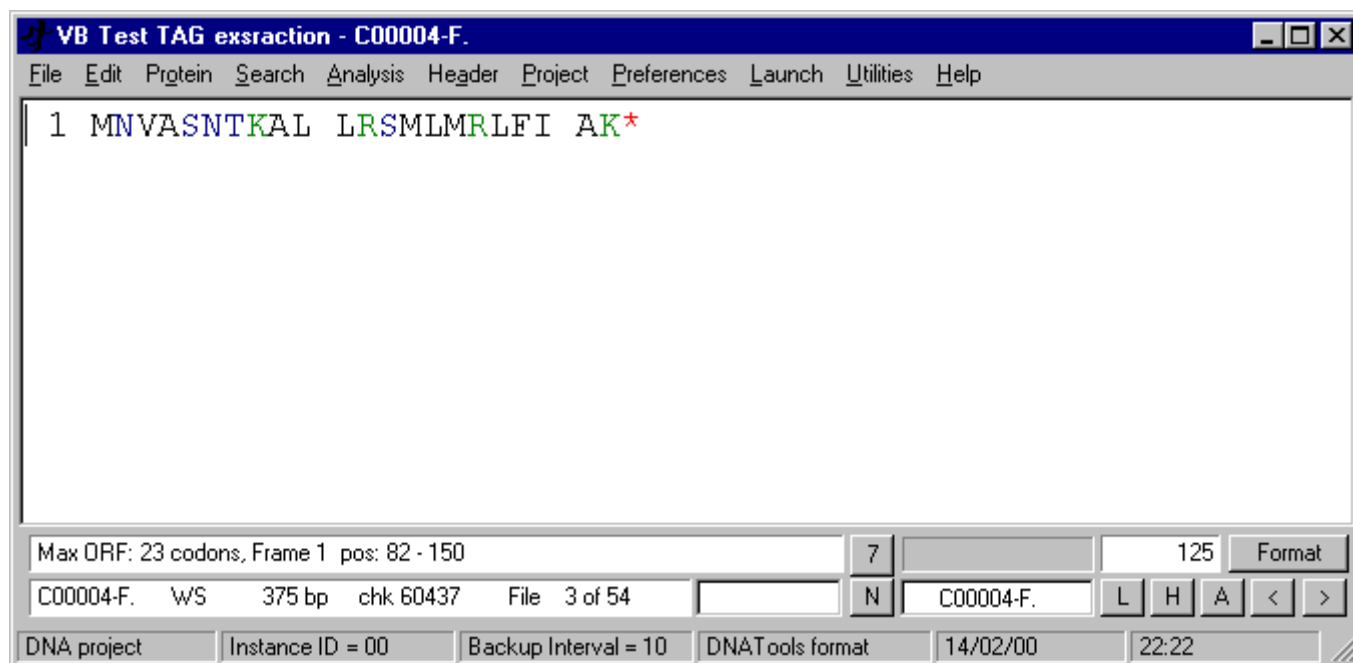


完全翻译, 一个读码框:

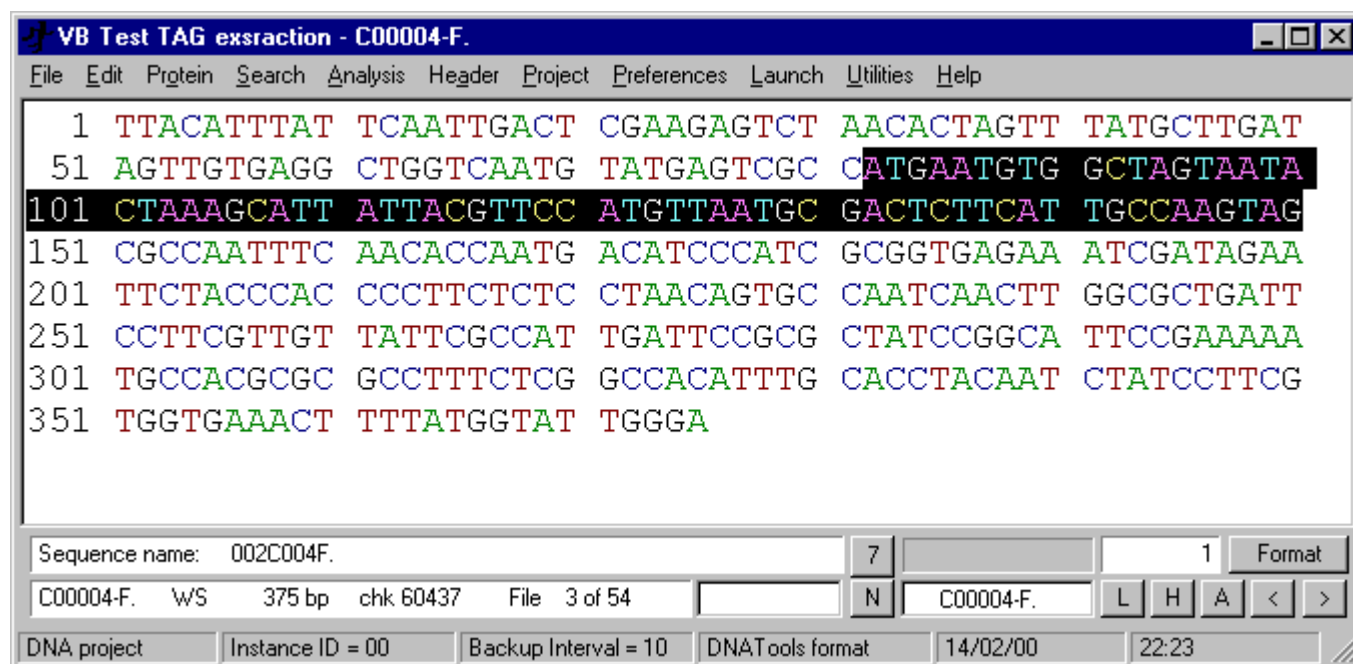
此功能“Protein/Complete Translation of One Frame”翻译完整的 DNA 序列(以指定的读码框)同时展示完整的包含终止密码子的蛋白质序列(以“*”标记)。翻译的序列可被保存以用于以后的数据库搜索使用。

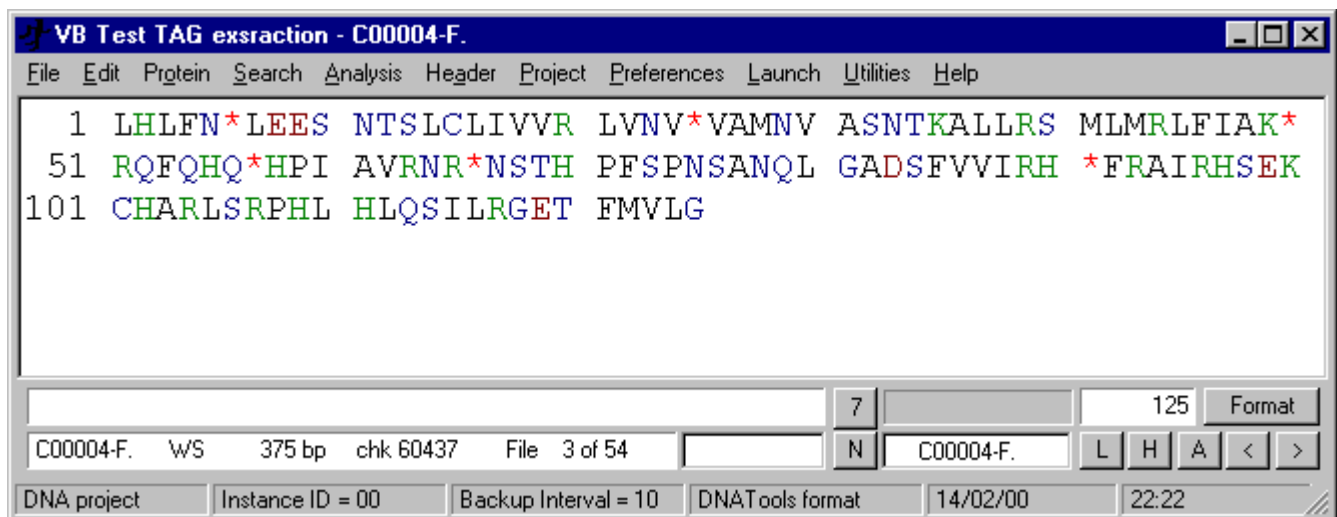
在一个读码框中寻找最长的 ORF:

此项功能“Protein/Largest ORF in One Frame”翻译 DNA 序列（以指定的读码框）同时展示最长的 ORF 和其相应的信息。所有的六个读码框的 ORFs 可以同时被展示，或者分开展示。



当翻译序列被展示时，点击格式按钮，高亮显示 DNA 序列中翻译的区域。



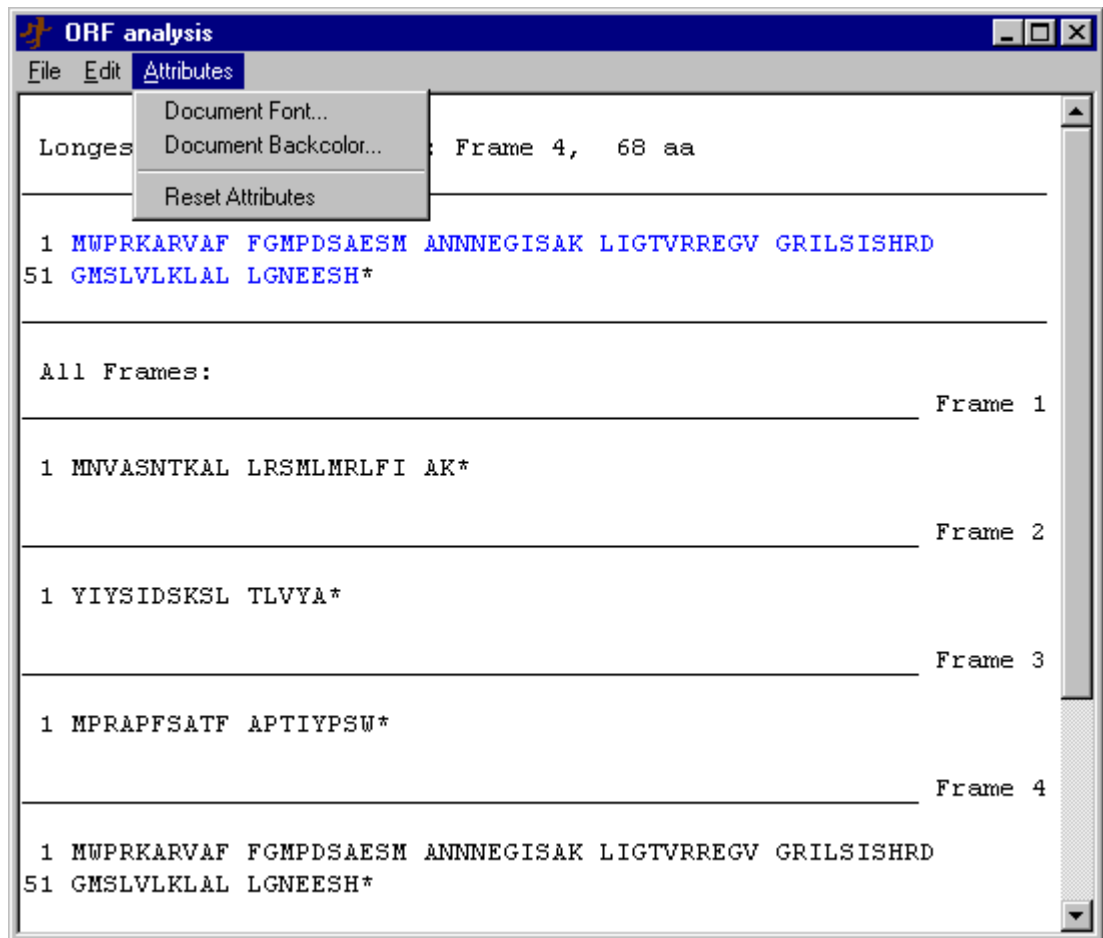


在一个读码框中寻找最长的片段：

此项功能“Protein/Largest Fragment in One Frame”翻译 DNA 序列（以指定的读码框）同时展示最长的未打断的蛋白质片段和其相应的信息。

在所有读码框中寻找最长 ORF：

此项功能“Protein/Largest ORF In Six Frames”翻译 DNA 序列同时展示最长的 ORF（每个读码框）。



在所有读码框中寻找最长片断：

此项功能 “Protein/Largest Fragment In Six Frames” 翻译 DNA 序列同时展示最长的氨基酸片断（每个读码框）。

评述：略

定义：

开放阅读框（ORF）— 在一个单阅读框中的一段 DNA 序列，按照以下限定：

DNA 序列的起始或者紧跟一个终止密码子的第一个 ATG 密码子；

DNA 序列的末端或者一个终止密码子（TAA, TGA, TAG）。

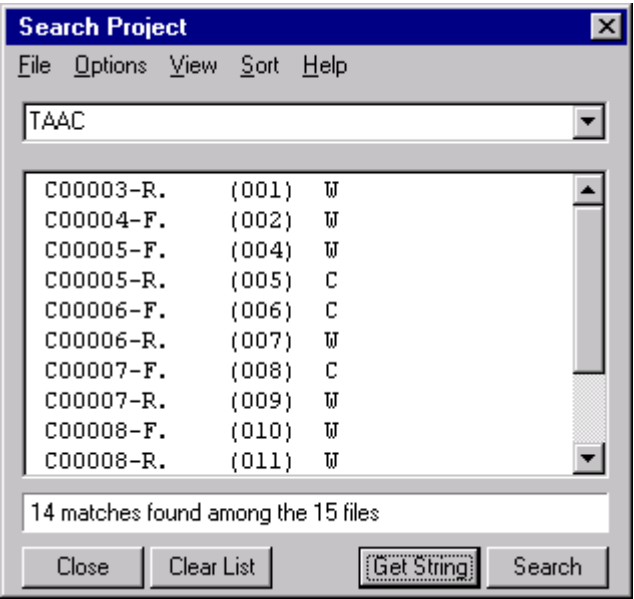
片断— 在一个单阅读框中的一段 DNA 序列，按照以下限定：

一个终止密码子或者蛋白质序列的起始；

一个终止密码子或者蛋白质序列的末端

3. 寻找方案 序列 标题：

此功能用于搜索系列标题（文本行），也可用于 DNA 或蛋白质序列（询问行）。在输入一个搜索行并点击搜索按钮之后，执行搜索，结果显示在列表框中。展示是受“View and the Sort”菜单中的设置控制的。



点击列表中的一行重新寻回序列或序列标题并高亮显示比对。高亮显示区域的内容接受用于序列搜索的“选项 options”菜单中的设置控制。

注意：此功能只寻找序列或标题中的询问行的第一个匹配。使用搜索序列功能以定位一个特定序列中的额外匹配。在文本标题帮助中有关于“搜索序列文本标题”的描述。

搜索模式（不适用于序列标题）

Symbols	Meaning
?	Any character.
[]	Any of the characters within the square brackets.
[!]	Any characters other than those within the square brackets.
5' /ABCn1-n2/	n1 to n2 chrs from 5' -end/N-terminal other than A, B, C.

/ABCn1-n2/3' n1 to n2 chrs from 3'-end/C-terminal other than A, B, C.
 /ABCcn1-n2/ n1 to n2 chrs other than A, B and C.

Examples:

Pattern:	Finds:	Does not find:
ASTS?V	ASTSxV	ASTSV
AST[GHWP]SV	ASTGSV and ASTHSV	ASTNSV and ASTRSV
AST[!GHWP]SV	ASTKSV and ASTYSV	ASTGSV and ASTHSV
/1-20/AST/4-8/SV	5' xxxASTxxxxSV	5' xxxASTxxxSV
AST/4-8/SV/2-20/	ASTxxxxSVxxx 3'	ASTxxxxSVx 3'
/A1-20/AST/4-8/SV	5' cbefASTacdefSV	5' cAefASTacbefSV
AST/B4-8/SV/1-20/	ASTacdeSVabc 3'	ASTaBdeSVabc 3'

X 是任何字符，5' 和 3' 分别代表 5' / N-终端 and 3' / C-终端。a, b, c 代表特定的字符。

提示，线索：

搜索行可以包含任何模式类型的组合；

若有或无被排除的字符的范围已经被指定适用于搜索行的 N 和 C 端，则 C 端范围被忽略掉；

为了在一个内部片断指定任何数量的字符，将低范围值设定为 0 并将上限值设定为序列的长度；

在范围内不允许的字符必须显示在指定范围值的前面；

为了使用包含终止密码子的行进行搜索(i.e. the character *), 包含星号在方括号中。

当没有包含在方括号中时，星号有与问号相同的作用；

常规的，当输入搜索行时，圆括号会被自动的转化为方括号；

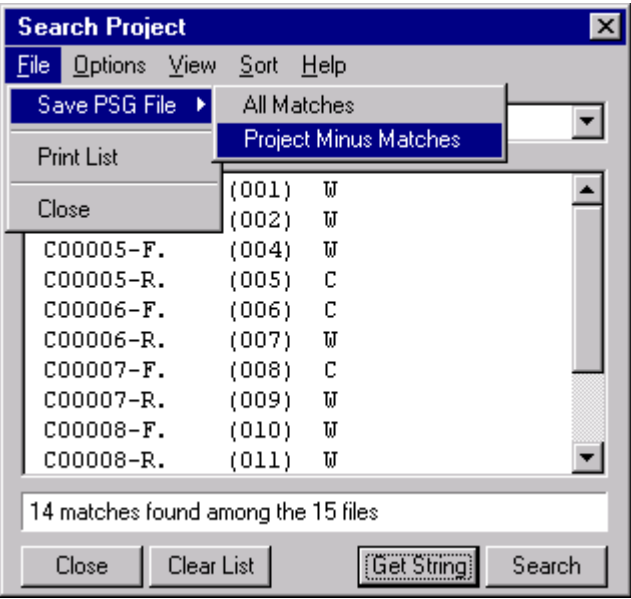
在搜索前，所有在搜索行中的字符被转化为上面的例子，例如：搜索总是例子不敏感的；

在启动搜索前，审查搜索行以找出不相配的斜线、括号和无效的范围。然而，这不完全排除以下可能性：即不能在一个序列中找到一个特定的模式是由于在搜索行中的语法错误；

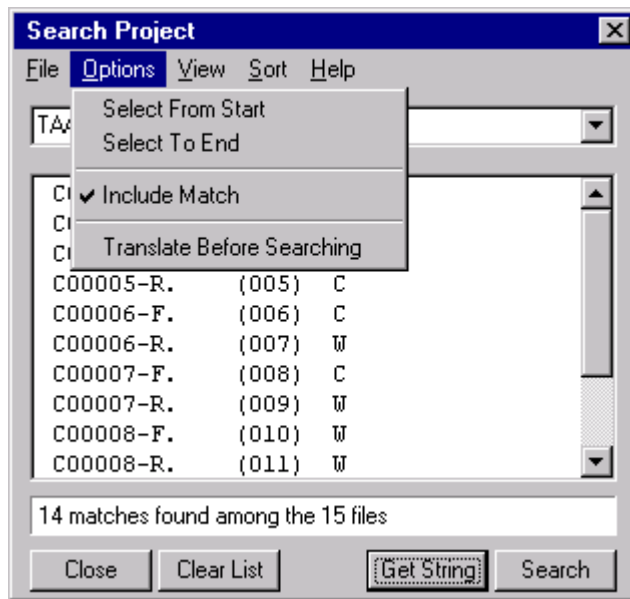
在搜索一个方案并评价匹配之后，程序推荐使用搜索序列功能重新针对每个比对进行搜索，看看是否序列包含超过一个匹配。

文件菜单：

保存为 PFP 文件一点击此选项保存在搜索中找到的序列（保存为 pfp 文件）。此 PFP 文件以后可用于打开一个新的只包含在搜索中找到的序列的方案。



选项（只在序列搜索中可见）



从启动中选择—高亮显示特定序列（从序列起始到匹配），用于移除序列的载体部分；

选择到末端—高亮显示特定序列（从匹配到序列的末端），用于移除序列的载体部分；

包含匹配—若上述的其中一个选项被核查，此设置决定搜索行是否包含在高亮显示的区域中；

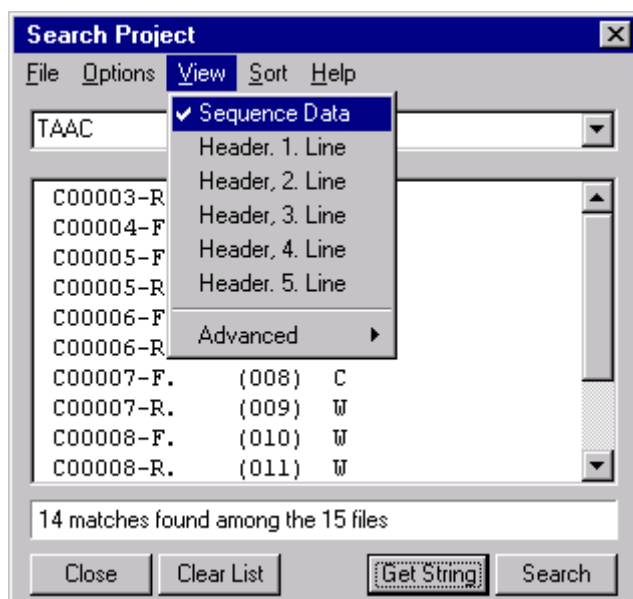
允许匹配中有终止密码子—若此选项被核查，程序将允许蛋白质序列中的匹配包含终止/起始密码子。否则这些匹配被拒绝。

在搜索前翻译—此选项允许用户在蛋白质水平搜索 DNA 序列。在搜索序列前，DNA 序列被翻译成蛋白质（以框架 1 起始）。翻译将以框架 2, 3, ... 6 继续进行，直到找到一个匹配（当然若存在的话）。当在结果列表中点击了匹配并且匹配区域被高亮显示，程序将重新寻回成功的框架。

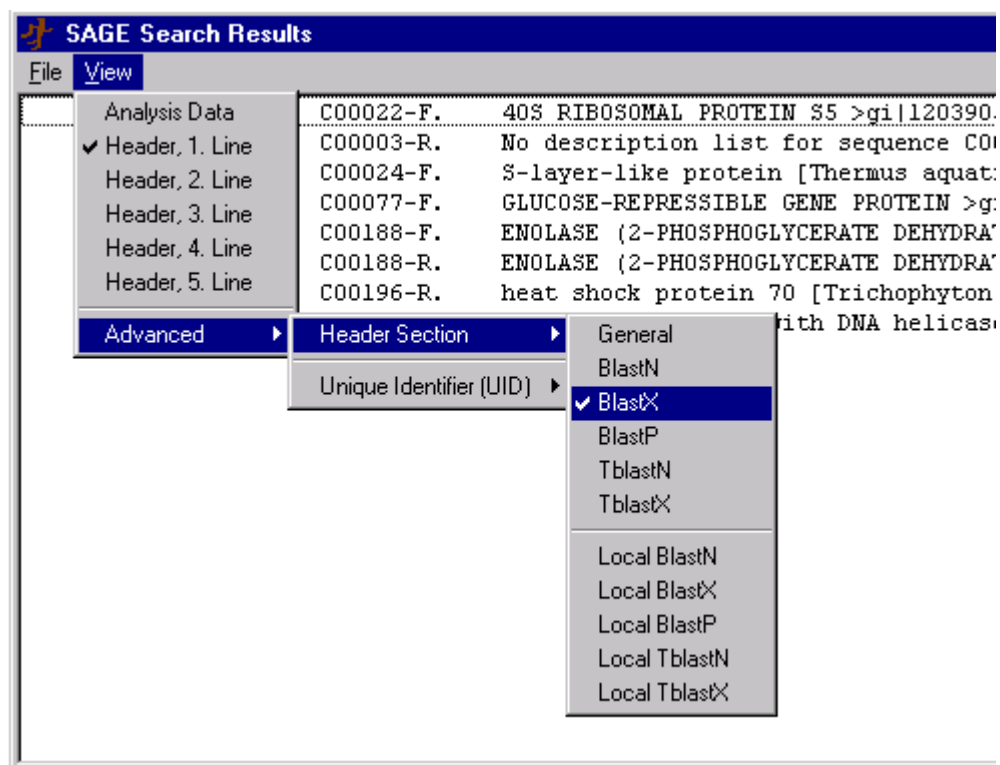
查看菜单：

序列数据—展示名字，长度，链（W or C）和匹配序列的核查总和。

标题行—展示匹配序列的标题行 1, 2, 3, 4 or 5



此表来自另外一个功能，但含“View”选项以用于改变序列在序列列表中展示方式。



分类菜单：

方案次序—依照他们被装载入方案中的模式（次序）来展示匹配；

按字母方式—按照字母方式分类结果列表。

4. 逆向翻译蛋白质序列:

Protein/Back-Translate Protein 命令执行将蛋白质序列翻译成退化的 DNA 序列且同时计算出退化程度。如果退化超过了 10000 次，真实值不被展示。

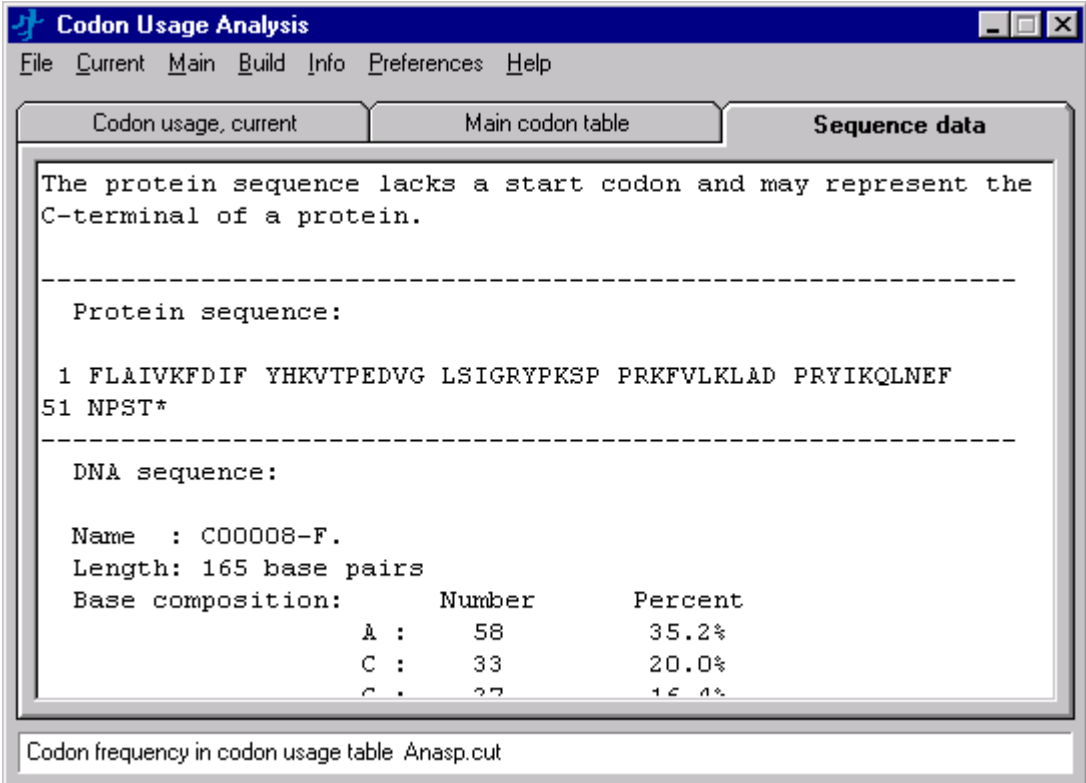
序列验证:

注意不要混淆含 GCG 特征的 DNA 序列和蛋白质序列。如果蛋白肽序列似乎是不正确的话（例如似乎是包含 GCG 特征的 DNA 序列），程序将提示警告。

自动编辑功能一万一回复翻译产生在序列 3' 端有一退化位点的 DNA 序列（如除了 M, ATG 和 W, TGG 之外所有的氨基酸），包含退化位点的翻译序列将被截断。

密码子使用表:

在回复翻译一个蛋白质序列之前，必须从文件菜单中选择密码子格式来装载密码子使用表。.



Codon Usage Analysis

File Current Main Build Info Preferences Help

Codon usage, current Main codon table Sequence data

The protein sequence lacks a start codon and may represent the C-terminal of a protein.

Protein sequence:

1 FLAIVKFDIF YHKVTPEDVG LSIGRYPKSP PRKFVLKLAD PRYIKQLNEF
51 NPST*

DNA sequence:

Name : C00008-F.
Length: 165 base pairs

Base composition:

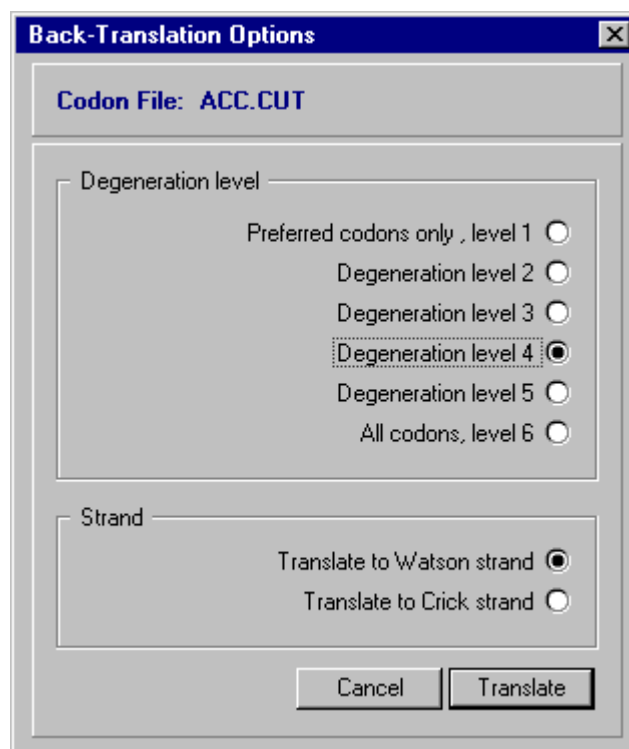
	Number	Percent
A :	58	35.2%
C :	33	20.0%
G :	37	22.4%

Codon frequency in codon usage table Anasp.cut

DNAtools 接受两种类型的表，用户生成的或用户修改的*.cut；或者是以 GCG 格式输入的特殊密码子使用表*.cod。后者可以被 DNAtools 简单的转为 DNAtools 的格式，即只需将其另存为*.cut 即可。然而不能保存密码子使用文件为 GCG 格式*.cod。

退化程度：

回复翻译的退化程度可以通过选择退化水平 1—6 进行控制，1 暗示只有首选的密码子才可以使用于回复翻译（结果的链是没有退化位点的）。选择 6 意味着可以获得最大退化，同时所有的可能性都包含在链中。2，3，4，5 意思是对每个氨基酸采用最经常使用的密码子。很明显，这将增加链的退化性，只要 2，3，4，5 对于一个给定的氨基酸有不同的密码子。



链：

蛋白质序列既可以被回复翻译成前向链也可被翻译成反向链。对于后者，在计算出退化程度之前回复翻译的序列就被转化为互补的链，同时移除不合规则的 3 端退化位点

5. 内含子编辑：

用于帮助用户定位基因组序列的内含子。此项功能确定序列中的所有的 GT-AG 对。

点击列表中的一行可以高亮显示当前展示序列中的区域或者以指定的框架(1, 2, 3 or all

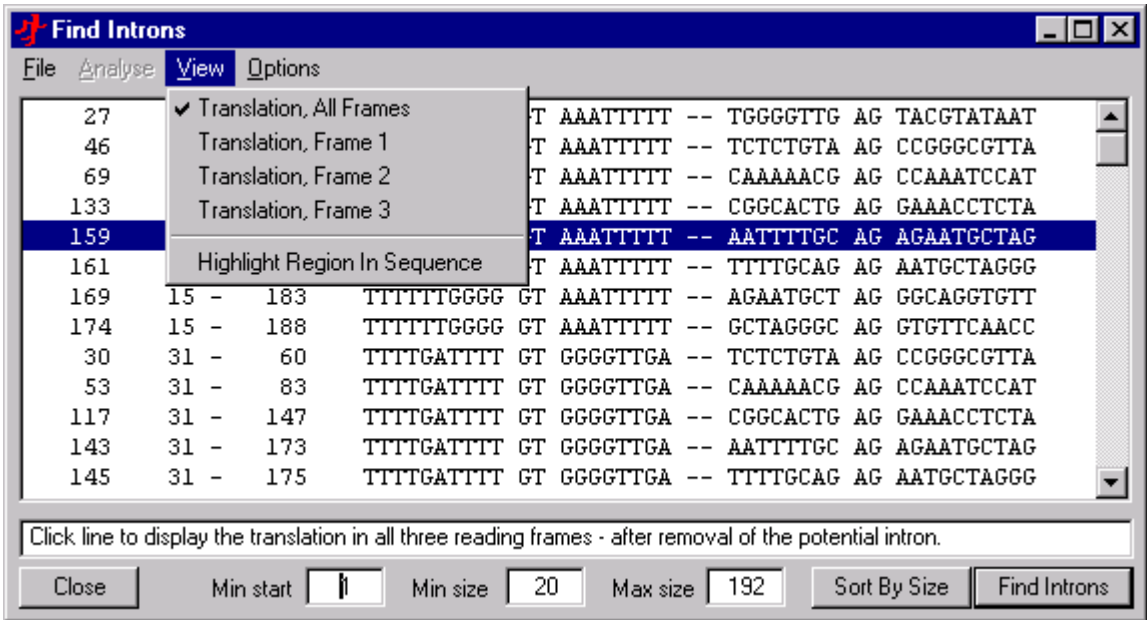
three) 翻译序列 (特殊的内含子移除)。若用户希望从序列中删除内含子, 只需点击<Delete> 键。

可以依据 GT-AG 对的起始位置来分类结果列表, 或者也可以依据内含子的长度来分类结果列表。

通过指定用于搜索的起始位置, 或通过设定内含子的最小和最大值来限制序列中确定的内含子/GT-AG 对的数值。

可以通过消除内含子 (这些内含子在整个序列的翻译[读码框 1, 2, 3]完成之后产生嵌在的可能的内含子列表。

注意: 此功能并不执行对内含子序列的任何评价, 仅用于定位和编辑内含子。



6. 自翻译 DNA 序列:

这个功能是为了帮助用户进行短 EST (表达序列标签) 序列分析而设计的, 以防通过数据库查询功能鉴定失败且正确的读码框又不知道。

这个功能以选定的读码框翻译所有当前方案中的 DNA 序列, 同时保存每个蛋白质序列为独立的文件。这些蛋白质序列可以针对蛋白质基序的数据库或者其他包含蛋白质信息的数据库进行查询。.

Create Protein Files

Translation options

Complete translation ☒

Longest fragment ☐

N-terminal fragment ☐

C-terminal fragment ☐

Frame options

All frames ☒

Forward frames, 1, 2, 3 ☐

Reverse frames, 4, 5, 6 ☐

Filter options

Minimum length

Format options

FastA file format ☒

GCG file format ☐

File Options

Each frame in one file ☐

All frames in one file ☒

Insert 5 x stop between frames ☒

Add two-char. frame-code to

Complete file name (8 chars.) ☒

Rightmost 6 chrs. of name ☐

Leftmost 6 chars. of name ☐

Middle 6 chars. of name ☐

Overwrite without warning ☒

C00003-R. _A11	742
C00004-F. _A11	748
C00004-R. _A11	771
C00005-F. _A11	769
C00005-R. _A11	661
C00006-F. _A11	697
C00006-R. _A11	770
C00007-F. _A11	873
C00007-R. _A11	762
C00008-F. _A11	802
C00008-R. _A11	946
C00009-F. _A11	872
C00009-R. _A11	769
C00010-F. _A11	883
C00010-R. _A11	772

The list includes 15 protein sequences longer than 60 aa

[Help](#) [Close](#) [Destination](#) [Create Files](#)

翻译选项：

完全序列：完全翻译的序列包含 X 和中止子；

最大片断：最大连续的氨基酸区（没有中止子）；

N-端区：以 M 开头且以第一个下游中止子为结尾；

C-端区：从序列开始到第一个中止子的区域。

框架选项：

所有的读码框；

前向 3 个读码框；（A, B, C）

反向 3 个读码框；（D, E, F）

过滤选项：

过滤选项允许用户忽略那些短于选定最小长度的蛋白质序列。

蛋白质文件名：

可以通过在 DNA 序列文件名中加_N 来定义蛋白质文件名。N 表示读码框（1—6 或者 # 表示在同一文件中所有的读码框）。这两个参数可以以下四种形式之一进行加入编辑。

用_N 替换 DNA 序列文件的扩展名；

将_N 加到文件名最左边的六个参数；

将_N 加到文件名最右边的六个参数；

将_N 加到文件名中间的六个参数。

在后面三种情况下，蛋白质文件名缺少扩展名。当建造蛋白质文件时，选定的文件名将被核实以避免相同的文件名。

如果选定的命名方法产生相同的文件名，创建过程将被停止同时建议选择另外一种产生蛋白质文件名的方法。万一四种方法都不能产生单一的蛋白质文件名时，初始的 DNA 序列文件名需要重新命名。

文件格式：

蛋白质文件可以以 Fasta 或 GCG 格式进行保存。每个蛋白质文件包含一个标题，给出序列名称，读码框和蛋白质序列的长度。蛋白质序列被打断成 50 参数的行，但没有行标记。

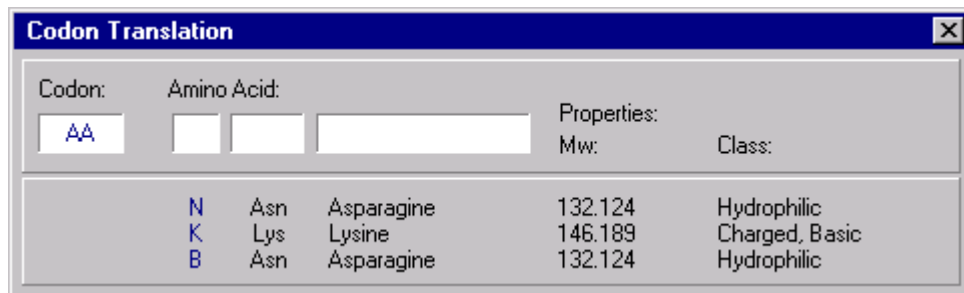
存储选项：

蛋白质文件既可以以独立的文件进行保存，还可以以每个 DNA 序列保存相应的蛋白质序列文件。如果选择后者且选择审查选项框时，在每个读码框之间将加入 5x 间隔物。

7. 密码子计算器:

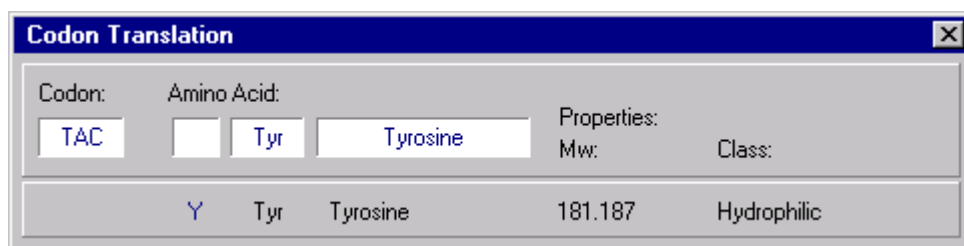
用于翻译密码子为氨基酸或反过来翻译。

对于某个特定氨基酸的有效的密码子列表包含相应的使用 GCG 特征表的退化密码子。Tab 键
键密码子和氨基酸输入栓牢在一起。当输入密码子时，N 是被允许的。



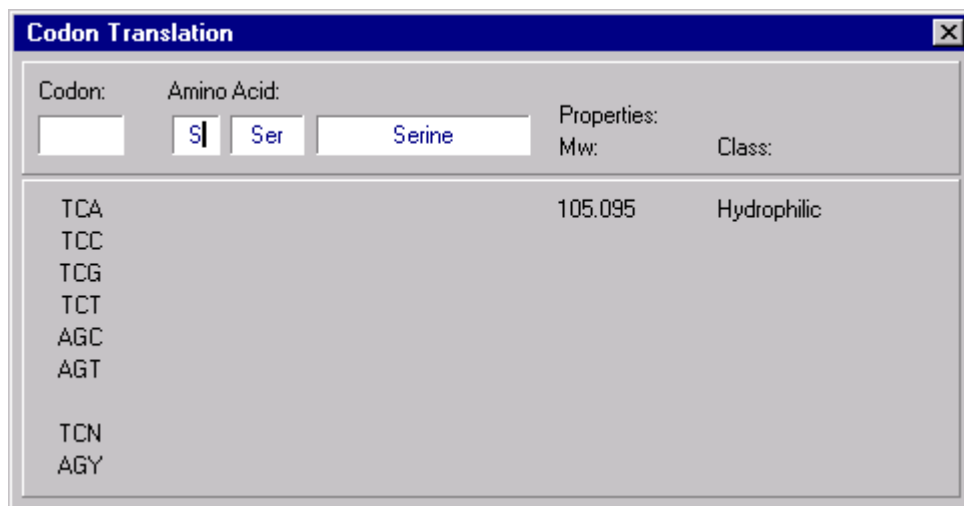
Codon Translation window showing the translation of Asparagine. The Codon field contains 'AA'. The Amino Acid field shows 'N', 'Asn', and 'Asparagine'. The Properties field shows 'Mw: 132.124' and 'Class: Hydrophilic'.

Codon	Amino Acid	Properties
AA	N Asn Asparagine	Mw: 132.124 Class: Hydrophilic



Codon Translation window showing the translation of Tyrosine. The Codon field contains 'TAC'. The Amino Acid field shows 'Y', 'Tyr', and 'Tyrosine'. The Properties field shows 'Mw: 181.187' and 'Class: Hydrophilic'.

Codon	Amino Acid	Properties
TAC	Y Tyr Tyrosine	Mw: 181.187 Class: Hydrophilic



Codon Translation window showing the translation of Serine. The Codon field is empty. The Amino Acid field shows 'S', 'Ser', and 'Serine'. The Properties field shows 'Mw: 105.095' and 'Class: Hydrophilic'.

Codon	Amino Acid	Properties
	S Ser Serine	Mw: 105.095 Class: Hydrophilic

Chapter7: DNAtools- trivial basic functions

1. 序列列表查看选项:

点击编辑器右下角的 L 按钮或者按下 CTRL+L。

列表包含旧的和新的名字，日期和文件最后保存的时间，序列起始和长度和一个序列特异的

5 数字审查总和。

接下来的密码是用于提示序列的起始：WS，沃森链；CS，克里克链；WI，反向沃森链；CI，反向克里克链。

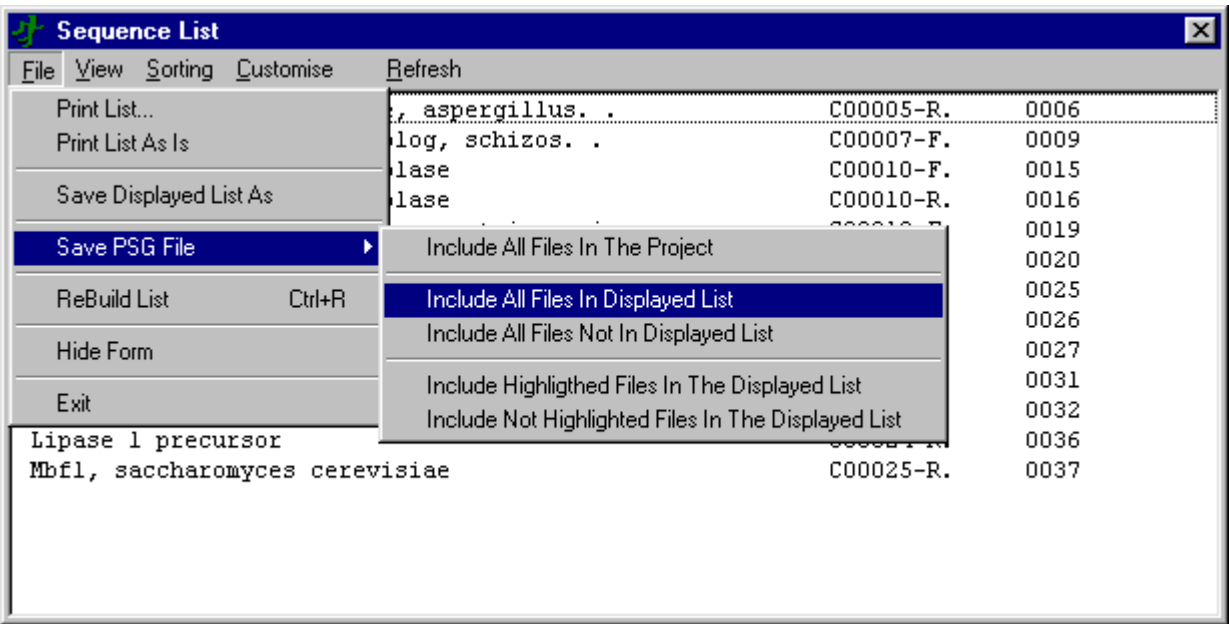
双击列表中的一个项目重新寻回编辑窗口中的选定序列。

文件菜单：

打印列表一点击该项打开打印表。选择完整的文件列表以打印序列数据或标题总结（如果用户希望打印序列标题的 1—5 行）。

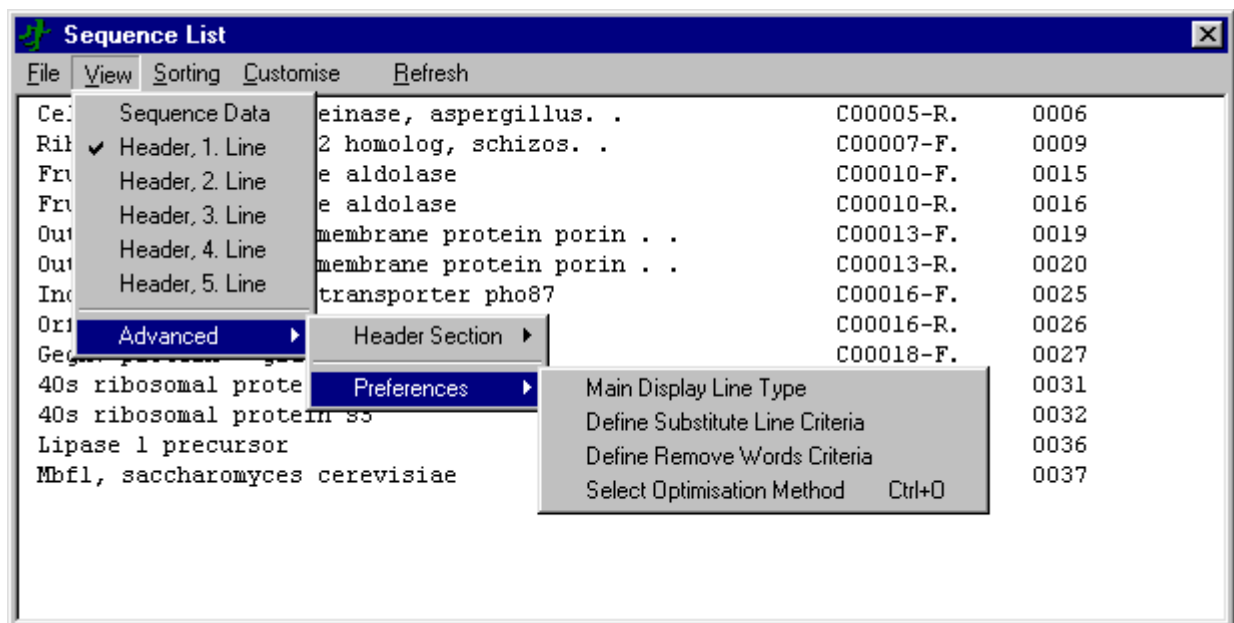
保存为一完整的序列列表可以被保存为一个正常的文本文件。这些文件还可以被输入到其他程序中。

保存 PSG 文件—允许用户保存序列文件的不同的亚组，依据不同的选项设置选定。



查看菜单：

允许用户选定哪种类型的信息用于生成序列列表。如果方案包含超过 700 个序列，用于展示序列列表的列表框容量将过超。 为了避免这些，序列列表的每一行被修短以容纳当前方案中的所有序列。



序列数据—通过状态行查看当前方案中的文件。

标题行 1—5—如果序列标题中的信息是和其他当前方案中所有文件中同一行的同一类型信息一起被系统的排列时，此选项是很有用的。使用 *General* 部分以包含使用者信息同时为同一类型的信息使用同一行。

For example:

- Line 1: Sequence name and origin
- Line 2: Highest homology in Blast search, DNA
- Line 3: Highest homology in Blast search, Protein
- Line 4: User comments

高级选项:

如果序列标题是 DNAtools 自动创建的并且包含 blast 数据库搜索结果、序列和或 Medline

记录，可以使用高级选项来列出序列文件（依据标题特定的日期）。用户可以选择依据一个特定部分（Blastn, Blast x, Blastp, Tblastn or Tblastx）的第一行来列出文件，含或不含 UID(unique record identifier)。

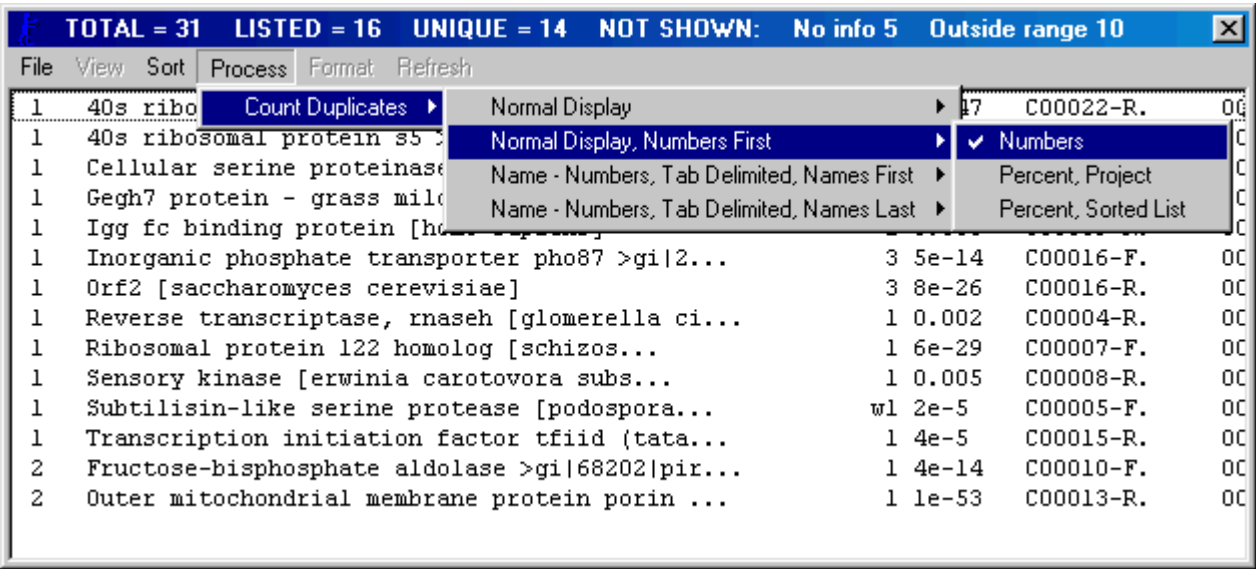
警告：如果用户没有按照自建格式来使用关于标题的高级列表选项，结果将不可预测。

分类菜单：

方案次序—以他们被装载入方案中的序列方式显示序列。

依字母顺序—显示当前方案中的以字母分类的序列。当序列列表是以标题行展示时，此选项尤其有用。这还可以激活加工选项。

加工菜单：



只有当分类序列列表被展示时，此菜单选项才可以用。此项功能基于 blast 搜索结果执行“clustering”，隐藏相同序列，只给出一个给定序列在方案中被找到的次数。这将给用户一个快速的对当前方案中的序列多余性的总结。注意此项功能只考虑名字并不执行任何形式的序列比较。

2. 转化序列为其互补序列：

转化为沃森链：此项功能“*Edit/Watson*”生成互补的 DNA 序列并显示它（其 5' 端在左边）；

转化为克里克链：此项功能“*Edit/Crick*”生成互补的 DNA 序列并显示它（其 5' 端在左边）；

颠倒序列：

Edit/Invert 序列功能颠倒当前序列，注意这个功能需谨慎使用。当拷贝那些以 3—5 方向写的序列时，这个功能很有效。在其他情况下，如颠倒那些 5—3 方向写的序列将导致与起始序列无联系。

注意：描述 DNA 序列起始的信息是和文件一起被保存的。当文件被装载时，这些信息可被寻回。在序列列表中，以下的密码提示序列的起始：WS，沃森链；CS，克里克链；WI，颠倒的沃森链；CI，颠倒的克里克链。

3. 无法打开网页：

4. 反向序列编号：

在主编辑窗口、显示翻译序列窗口和显示限制性图谱窗口中逆转序列的编号。

Before:

1 ATGCTAGATGATAGATAGAT

21 ATGCTAGATGATAGATAGAT

After:

40 ATGCTAGATGATAGATAGAT

20 ATGCTAGATGATAGATAGAT

5. 弥补序列编号：

当前序列的编号可以被偏移为一个位点随意编号。当序列与序列图比较时，可以弥补移除的 5' 端。用修短的序列的 5' 部分的长度抵消数字，序列的编号可以与序列图的一致。

Before:

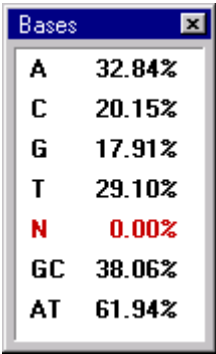
1 ATGCTAGATGATAGATAGAT
21 ATGCTAGATGATAGATAGAT

After offsetting by 100:

101 ATGCTAGATGATAGATAGAT
121 ATGCTAGATGATAGATAGAT

6. 序列基本组成:

显示碱基组成 (AT, GC 和 N)。该 “*Analysis/Base Composition*” 表呆在编辑器的上方。



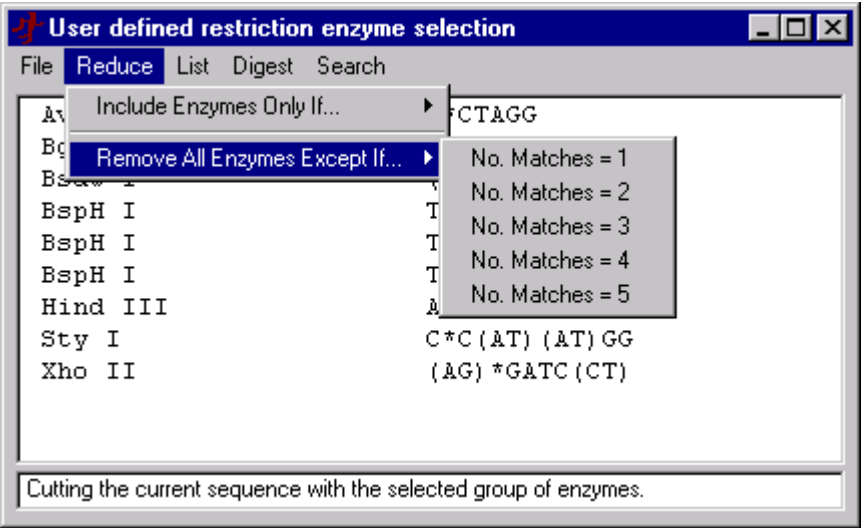
A	32.84%
C	20.15%
G	17.91%
T	29.10%
N	0.00%
GC	38.06%
AT	61.94%

7. 限制性图:

此项功能 “*Search/Start Search* 或者 F3” 被用于搜索当前展示于编辑窗口中的序列（使用一组限制性位点或用户定义的搜索队列）。如果序列的一部分被阻断，只有那些被阻断的部分才被搜索。对于非回文序列的搜索准则，序列的两条链都被搜索。

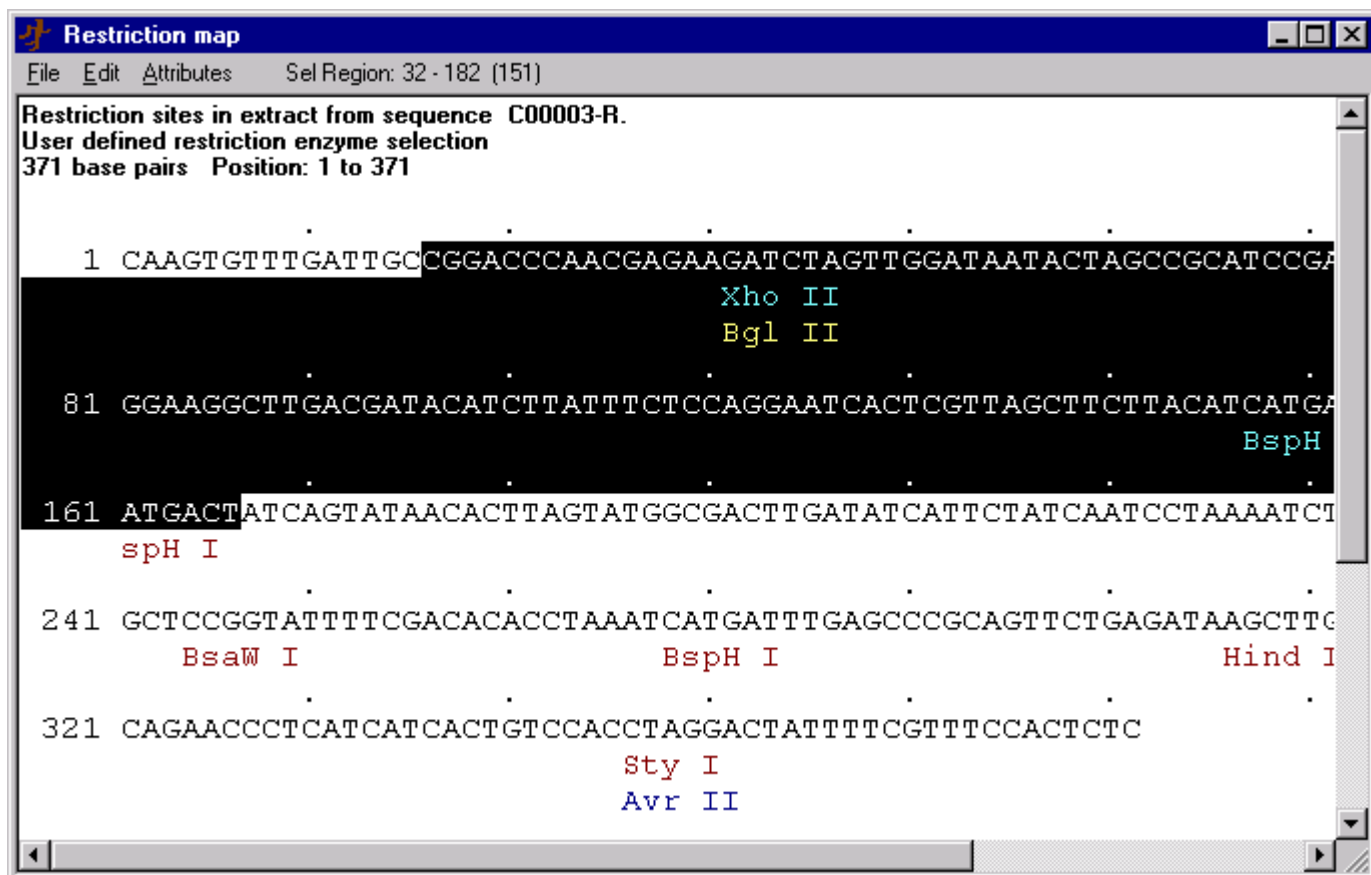
Hits/质粒编辑器列表:

此 “*Search/Start Search/Plasmid Editor*” 表显示搜索结果并包括匹配搜索队列的序列和名字和其在 DNA 序列中的相配之物。剪切位点一万一时限制性位点一用信号注释。此表中的数据可以帮助用户设计克隆试验。



序列中的 hits 图（限制性图谱）：

使用 “*Search/Start Search/Show Map Of Hits*” 选项，搜索结果可以以完整的序列被展示，或者作为序列的阻断部分和限制性位点的名字一起展示。图中搜索队列名字的第一个字符代表限制性酶切点最右边的碱基，或者代表用户定义的搜索队列的第一个碱基（无剪切位点）。（基序、引物）



注意：在图谱中可以展示九个完全或部分重叠位点（最大值）。如果不是所有的位点都被显示，程序将通知用户。为了避免这些问题，用一个更严格的限制性位点选择方式重复搜索。

在选定的组中的限制性酶（并不剪切特定的序列或序列提取）列表显示在下图中。

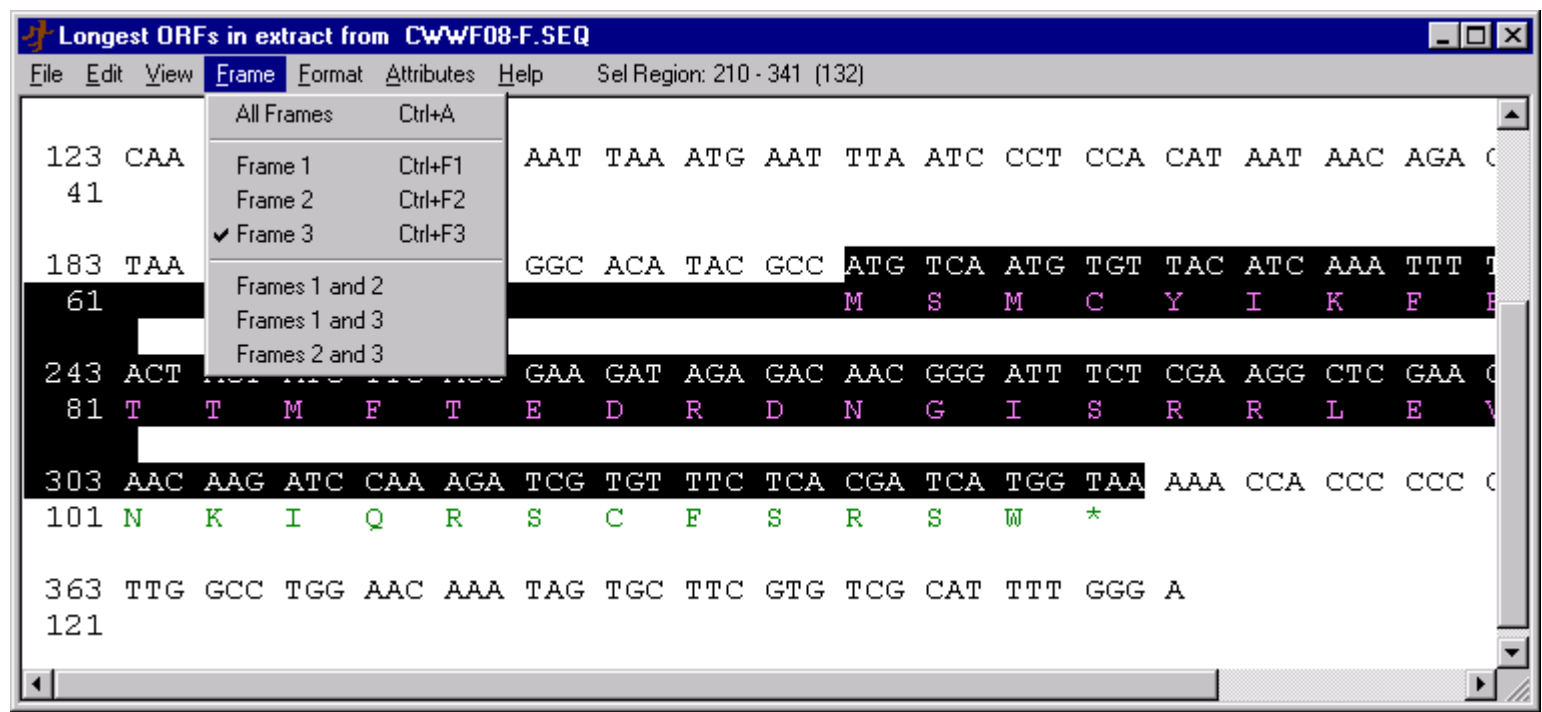
锚—可以使用此选项选择一个序列区域（基于限制性图谱）。当打开结果表时，锚 1 被激活用于输入该区域的底端界限。在点击选定碱基的右边之后，锚 2 被选定（或者按下 CTRL+F2 或者从主菜单中选择）。再次点击碱基的右边。点击 Pos 菜单项目或者从组菜单中选择“Anchor/Show Region”。这将展示选定的区域。当区域仍是高亮显示时，关闭表以高亮显示展示在主编辑器中的序列区域。这个区域可以被取代，拷贝或删除。

位点—展示当前选定的锚和锚所定义的区域长度。

无切割—酶列表（不剪切所分析的序列）可在限制性图谱的底端找到。

8. 翻译 DNA 序列：

此 “*Protein/Translate Sequence Extracts*” 表展示当前 DNA 序列的翻译情况。



行数相应于提取序列区域的相配之物。终止密码子用星号注释，用 X’ 注释不确定氨基酸。

DNA 序列的格式与序列编辑表中选定的格式无关（阻断长度是 3，行长度 60bp）。

文件：

打印—打印序列和翻译（就像显示于屏幕上的一样）。

查看：

完整翻译—展示 DNA 序列完整的翻译（以选定的读码框）。

所有的开放式读码框—展示所有开放式读码框（以选定的读码框）。

最长开放式读码框—展示最长的开放式读码框（以选定的读码框）。

最长片断—展示两个终止密码子之间最长的片断（以选定的读码框）。使用此选项定位序列错误（产生读码框漂移错误）。

框架:

所有框架一在 DNA 序列之下展示三个前向读码框的氨基酸序列。

框架一翻译 DNA 序列并在 DNA 序列之下展示氨基酸序列（以选定的前向读码框）。

高亮显示一个区域:

使用锚一使用此选项，可以在氨基酸序列的基础上选择一个序列区域（如在一个开放式读码框中的一个内含子）。当表打开时，锚 1 被激活用于输入区域的底端限制。在点击选定碱基的右边之后，锚 2 被选定（或者按下 CTRL+F2 或者从主菜单中选择）。再次点击碱基的右边。点击 Pos 菜单项目或者从组菜单中选择“Anchor/Show Region”。这将展示选定的区域。当区域仍是高亮显示时，关闭表以高亮显示展示在主编辑器中的序列区域。这个区域可以被取代，拷贝或删除。

通过拖动一高亮显示序列的一部分。*Mouse-Click -> Mouse-Down -> Mouse-Drag -> Mouse-Up* 有相同的作用: 当一个区域被高亮显示时关闭翻译表，高亮显示的部分被维持显示（当序列展示在常规序列编辑器中）。在开始拖动之前，务必抓住区域的起始部分。选定的起始位点展示在主菜单中。当区域仍然被高亮显示时，关闭表以高亮显示展示在主编辑器中的序列中的区域。区域可被拷贝、删除。

位点:

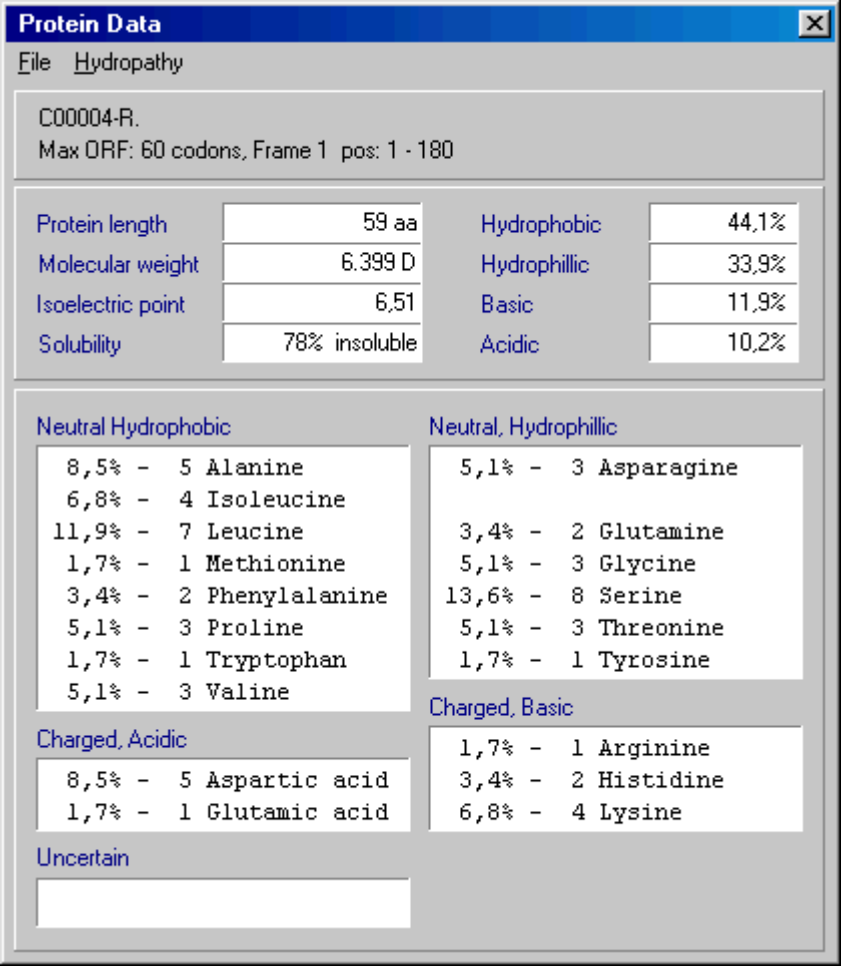
展示当前选定的锚/起始点和锚所定义的区域长度。

Chapter8: DNAtools- protein analysis

1. 蛋白质属性 分子量 等电点 氨基酸:

该“*Analysis/Protein Properties*”表显示一个关于蛋白质序列的简单描述，包括长度，分子量，氨基酸组成以及带电和疏水性氨基酸的分布。

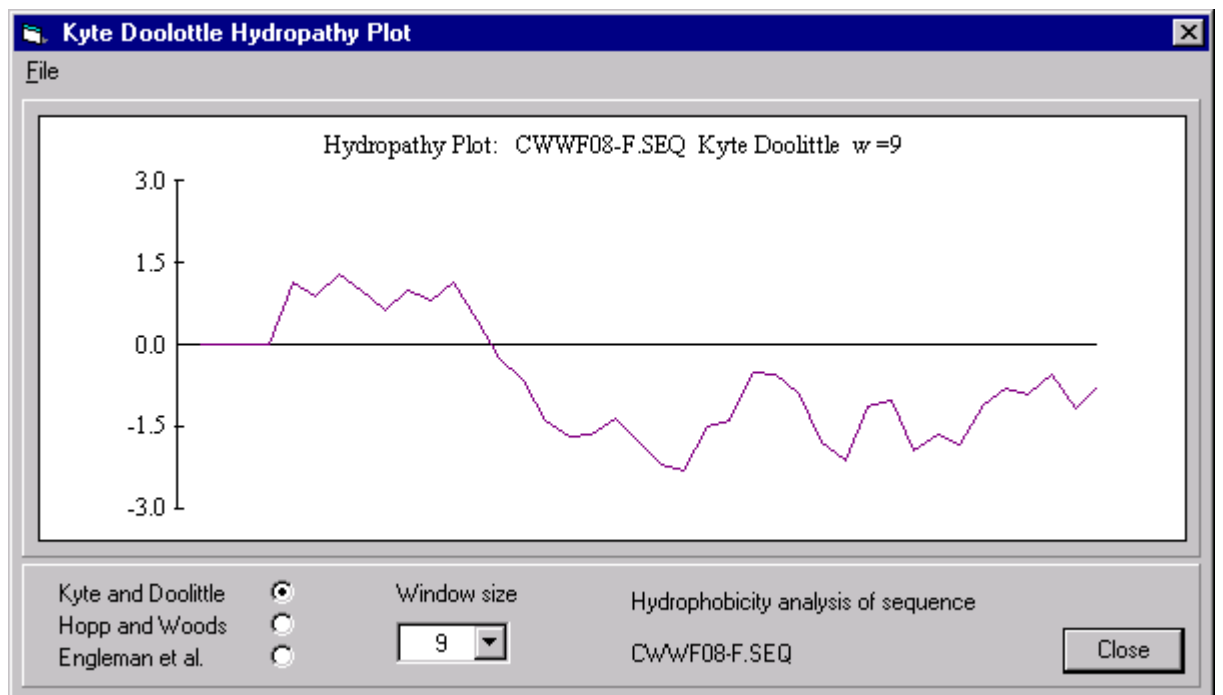
通过修订的 Wilkinson-Harrison 溶解性模型和两个参数模型（Davis, G. D. and Harrison, R. G. (1999)）来计算溶解性。新的融合蛋白系统用于 *Escherichia coli* 的可溶性表达。Biotechnology and Bioengineering 65, 382 – 388.



2. 蛋白质的疏水性:

此项功能逐渐的用于估计蛋白质的亲水性和疏水性以及其氨基酸序列。移动窗口的大小可在 3 到 19 个氨基酸残基之间变动。此功能计算移动窗口中氨基酸的平均疏水性值。有三个数据集可以获得: Kyte and Doolittle, Hopp and Woods and Engleman 等。

当任何一个参数发生改变时, 此功能立即重新计算出疏水性情况。使得评价改变窗口大小和数据集所带来的效果变得很容易。



References:参考文献

Kyte, J, and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. J, Mol. Biol., 157, 105-132

Engleman, D.M., Steitz, T.A. and Goldman, A. (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Ann. Rev. Biophys. Biophys. Chem., 15, 321-353

Hopp, T.P. and Woods, K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. USA. 78, 3824-3828

3. 密码子计算器:

用于翻译密码子为氨基酸或反过来翻译。

对于某个特定氨基酸的有效的密码子列表包含相应的使用 GCG 特征表的退化密码子。Tab 键键密码子和氨基酸输入栓牢在一起。当输入密码子时，N 是被允许的。

Codon Translation					
Codon:	Amino Acid:			Properties:	
AA				Mw:	Class:
	N	Asn	Asparagine	132.124	Hydrophilic
	K	Lys	Lysine	146.189	Charged, Basic
	B	Asn	Asparagine	132.124	Hydrophilic

Codon Translation					
Codon:	Amino Acid:			Properties:	
TAC		Tyr	Tyrosine	Mw:	Class:
	Y	Tyr	Tyrosine	181.187	Hydrophilic

Codon Translation					
Codon:	Amino Acid:			Properties:	
	S	Ser	Serine	Mw:	Class:
TCA				105.095	Hydrophilic
TCC					
TCG					
TCT					
AGC					
AGT					
TCN					
AGY					

Chapter9: DNAtools- multi-sequence functions

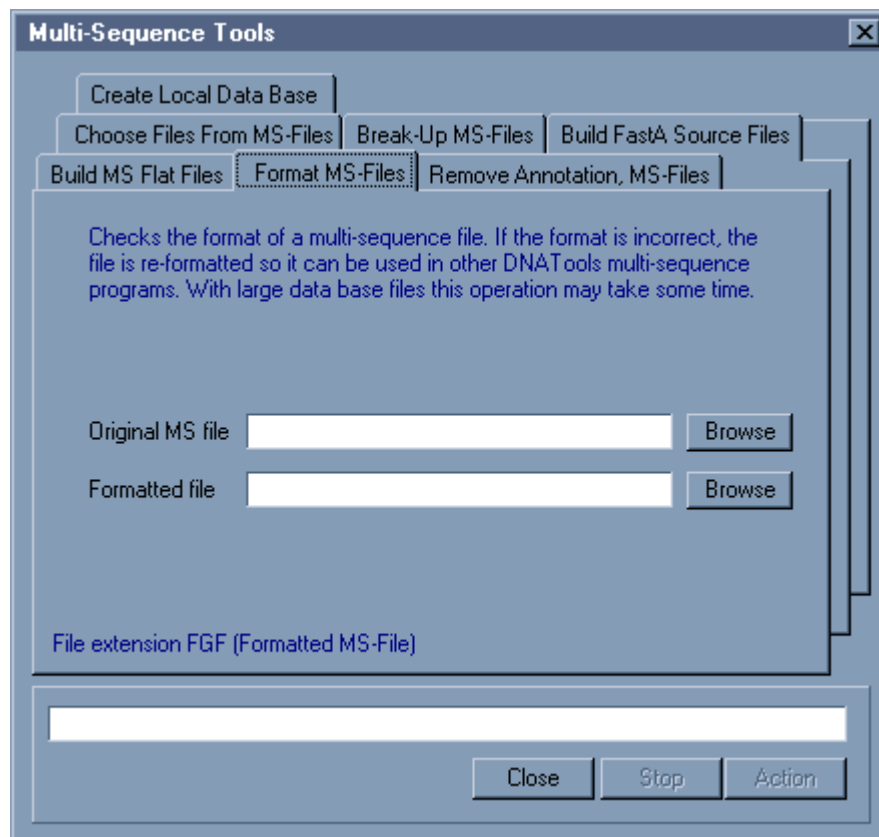
1. 多序列工具:

此项功能允许用户修改或生成多序列文件。FlatFile 或 FastA 格式的数据库文件包含文本文件，而这些文本文件又含纯 ASCII 格式的多数据库且此数据库文件可从几个 WEB 站点如 NCBI 处下载到。(ftp://ncbi.nlm.nih.gov/genbank/)

格式:

此项功能审查多序列文件的 EOL(end of line)码以确认 LF 是否单独使用。如果的确如此，所有的 LF 码被 CRLF 码代替且格式化后的文件以相同的名字被保存，但扩展名改为

.fgf (Formatted Genbank File). 万一正确的 CR 码在被使用，该文件的一个拷贝以扩展名为. fgf 被保存。



移除注释：

此项功能移除来自 flatfile 格式的 GenBank 文件中序列的注释，但保留 DESCRIPTION 和 ACCESSION 行不变。初始的分离子和//记录分离物在修改的文件中仍然被保存下来。

如果选择审查 “Exclude sequences > 15,000 bases”，那么长于 15000 碱基的序列被从修剪文件中排除掉。在大多数情况下，如此长的序列包含多基因序列，从此序列中提取的 SAGE 标记是无意义的。

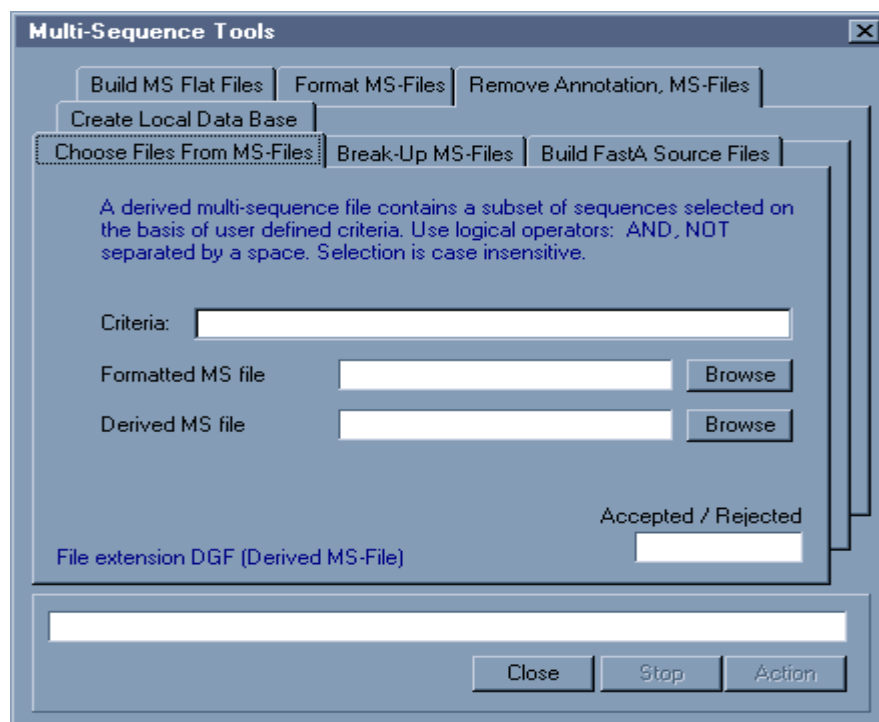
输入文件必须是扩展名为*.fgf 或 *.ngf，如用户生成文件或一个 GenBank 文件且这些文件的 EOL 码已经被审查并且若不正确则被替换。修改的文件以相同于输入文件的文件名被保存，但扩展名改为*.tgf。(Trimmed GenBank File)。

衍生 GenBank：

此项功能允许用户从格式化的 GenBank 文件中提取一亚组序列。如从完整的 EST 数据库平台文件中提取一个特殊生物的 EST 序列。GenBank 文件必须是标准的格式即:每个记录的序列部分被 ORIGIN 和 //所划分。注释部分必须包含 DEFINITION , ACCESSION, ORGANISM 和 REFERENCE 部分。对于 FastA 格式的多序列文件或修改过的 GenBank 文件不起作用。

标准: 用于为衍生的 GenBank 文件选择文件的准则—使用逻辑算法 AND 和 NOT 输入, 例子如下:

- AND Homo (retains all files containing the word Homo/HOMO/hoMo in the annotation)
- NOT Yeast NOT Saccharomy NOT cerevisiae (selects all sequences except yeast sequences)
- AND Plant (retains all plant sequences)
- AND Fungi (retains fungal sequences)



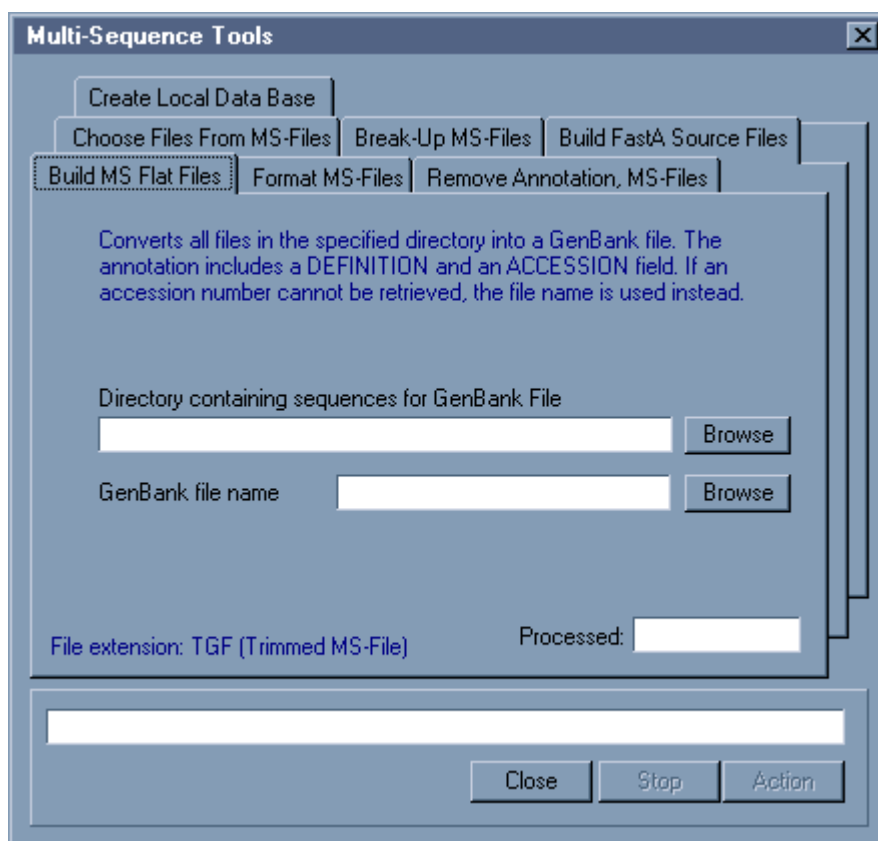
注意: 在标准中, 最大值只允许 5 个 AND 和 5 个 NOT。若每个类别中超过了 5 个, 在选择时这些过多的词被忽略掉。逻辑算子和关键词必须用空格隔开。搜索是案例敏感的。

输入文件必须是扩展名为*.fgf 或 *.ngf，如用户生成文件或一个 GenBank 文件且这些文件的 EOL 码已经被审查并且若不正确则被替换。修改的文件以相同于输入文件的文件名被保存，但扩展名改为*.tgf。(Trimmed GenBank File)。

明显的，此项功能只与 GenBank 文件相关，此时注释没有被修改。

生成新的：

用于依据自己的序列生成一个新的 flatfile 格式的 GenBank 文件。在尝试创造文件之前，确信所有包含于文件中序列定位于相同的目录。新的 GenBank 文件的记录格式与那些修改过的 GenBank 文件相同而且新文件的扩展名为*.ngf (New GenBank File)。



唯一保留的序列鉴定信息是文件名，在 DESCRIPTION 后被包括在内。所有包含于标题中的其他信息则被丢失。

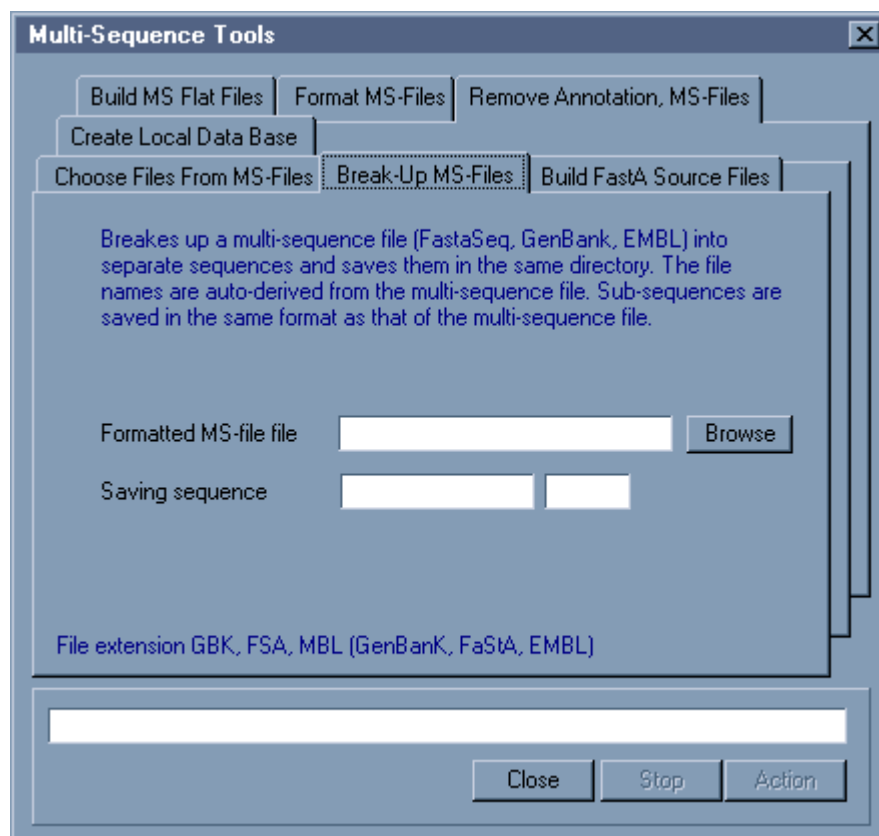
DNAtools 接受 GenBank, DNATools 和 FastA 格式的文件。在某些场合, 当整个数据库文件出现在文件标题中时, 大于一个标题/序列分割子将会出现在文件中。这将导致标题/注释和序列不正确的分离。

解决此问题的方法是装载文件入 DNAtools 中并且再一次保存他们: 在文件被保存前, DNAtools 审查每个文件的标题, 看看是否有不合规定的分离子。如果有多个分离子出现在标题中, 他们将会被转化为分离子不认识的参数。

此项功能可以被用于生成一个多序列文件以用于 SAGE 标签提取。换句话说, 用户可以装载大量的文件入方案中, 相反的, 首先将其合并成一个多序列文件然后使用 SAGE 提取功能以生成 SAGE 标签文件。

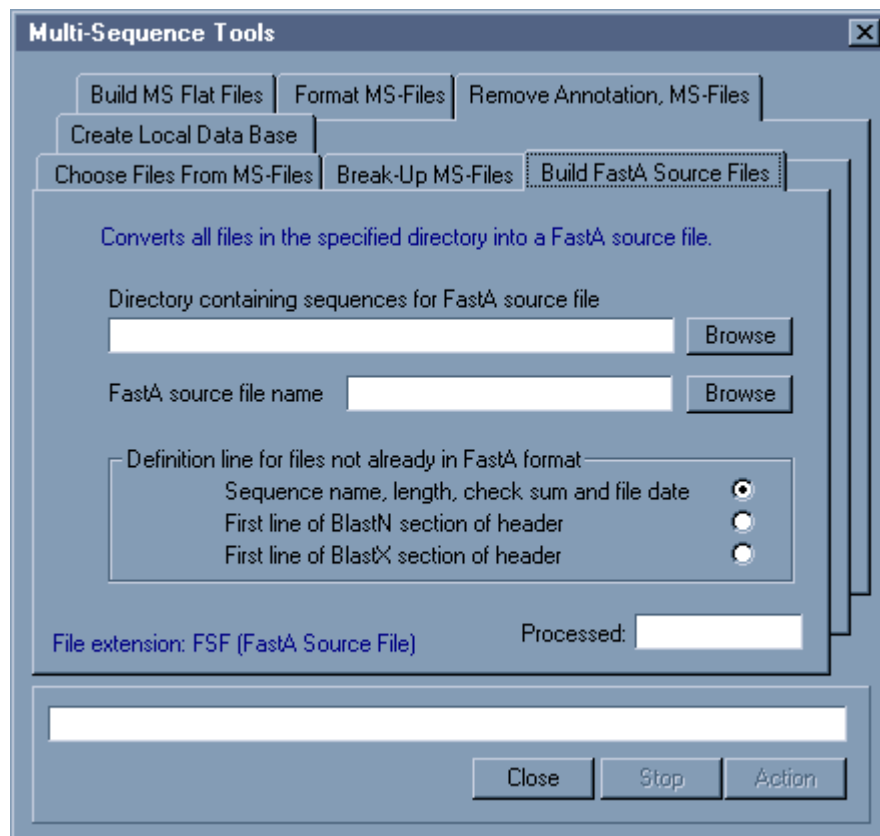
打碎多序列文件:

此项功能打碎一个 GenBank 或者 FastA 格式的多序列文件同时在相同的目录下保存每个子文件。子序列的文件名由这些子序列的索取号组成, 扩展名为*.gbk 的是用于 GenBank 文件而扩展名为*.fsa 是用于 FastA 文件。



生成 FastA 源文件：

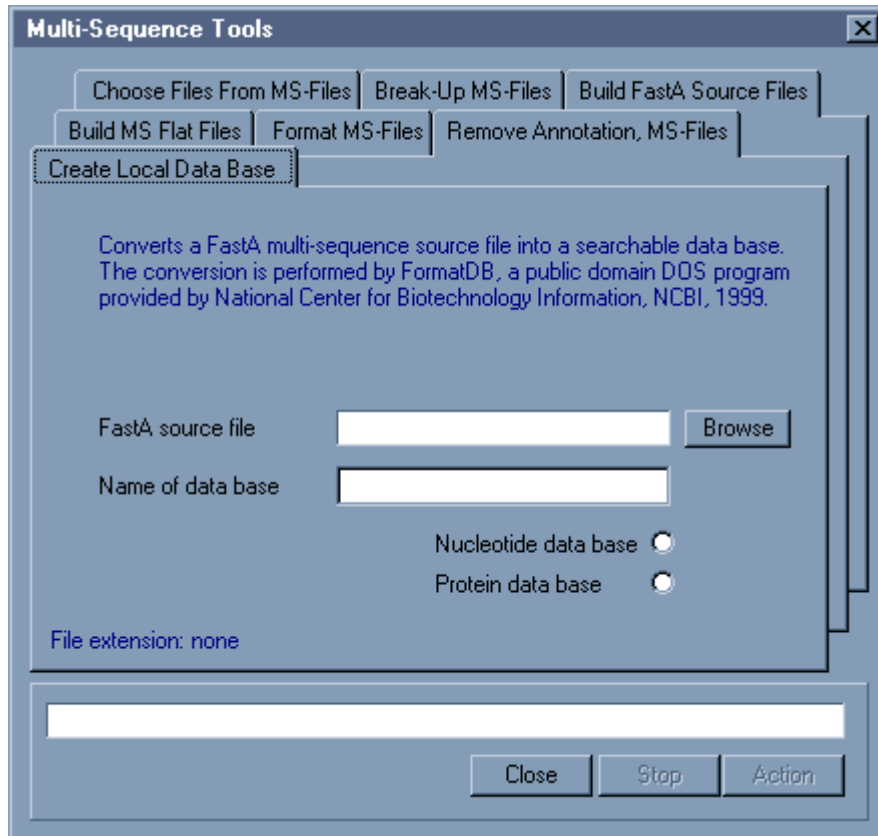
为了生成一个 FastA 源文件，有必要收集所有的特定序列（即用户期望这些序列包含在同一目录下的源文件中）。如果 DNAtools 已经将这些序列进行注释，用户有必要提示用户希望哪个标题部分 (blastN or blastX) 用于注释这个源文件。



生成本地数据库 (with formatdb, NCBI)：

此项功能运用 NCBI 的 DOS 程序 FormatDB 以生成可搜索的数据库。

查看 “*Local Blast Search*” 寻找 formatDB DOS 程序的安装信息。为了生成一个数据库，按照以下步骤：



如何生成一个本地数据库：

拷贝所有的用户希望将其包含在本地数据库中的序列到一个空的目录下；

点击 “Utilities/Multi-Sequence Functions/Build FastA Source” 从目录中的文件以生成一个 FastA 源文件；

如果序列文件是 DNAtools 格式，选择注释源文件-否则注释将自动的从源文件中被提取出来；

点击 “Utilities/Multi-Sequence Functions/Create Local Data Base” 以建造本地数据库；

记住：如果正确的数据库类型被选择了，审查核实一下；

完成的本地数据库被创建在 Windows / Winnt 目录下的 DT5_TEMP 子目录中，同时如果希望搜索它时必须将其保留在那里；

万一希望从含 DNA 序列的方案中生成一个蛋白质数据库，在建造 FastA 源文件和最终的数据库之前，使用“Utilities/Create protein files”翻译该核苷酸序列。

注释：该数据库中序列将只包含一行注释。如果输入序列是 GenBank, FastA 或 GCG 格式，注释将被自动的从初始文件中寻回。对于 GenBank 文件，使用 DESCRIPTION 行；对于 FastA，使用标准的单行注释；对于 GCG，.. 分割符号之前跟行。

对于 DNAtools 文件，列在表上的标题部分选择哪个部分用于注释。默认的是序列名，长度，审查总和和文件日期。如果文件标题包含 blastn 和或 blastx 搜索结果，其中的一个标题部分的第一 DESCRIPTION “描述” 行可被选择用于数据库入口的注释。对于一个或多个序列标题，如果选择的标题部分 (Blastn: or Blastx:) 是缺失的，将使用默认的设置。

万一用户的序列包含一个 Blastn 和一个 Blastx 标题部分，用户可以用 blastn 注释生成一个数据库并且可以用 blastx 注释生成第二个数据库。

2. 生成本地数据库：

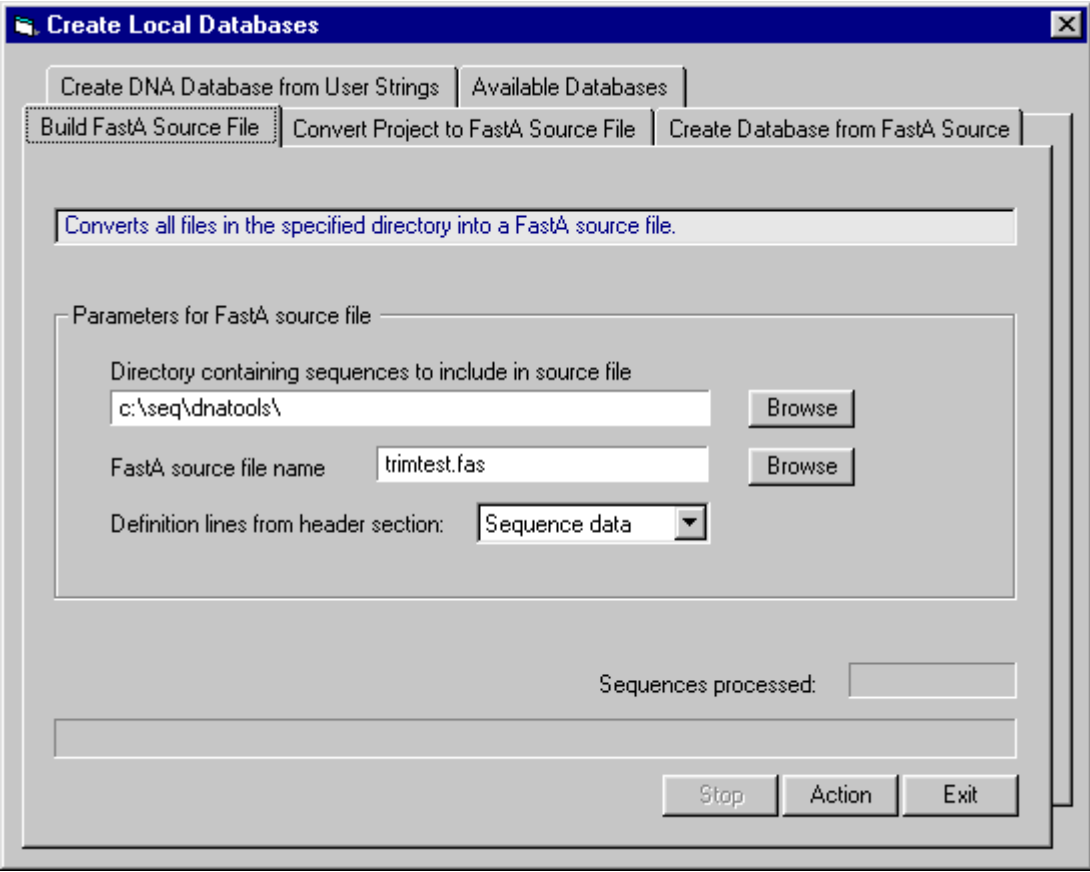
生成本地数据库文件：

这张表包含几个功能用于生成本地数据库文件。生成本地数据库包括两步： 1，为 NCBI 员工提供的 formatdb 程序创建一个源文件；2，输入这些源文件并运行 formatdb。程序收集允许用户合并一个特殊目录中所有的文件、所有 DNAtools 方案中的序列和所有自己输入到源文件中的序列。这张表同时包含一个简单的管理工具，当用户想要从列表中移除数据库时，这个工具是很方便的。

从目录中的文件生成源文件：

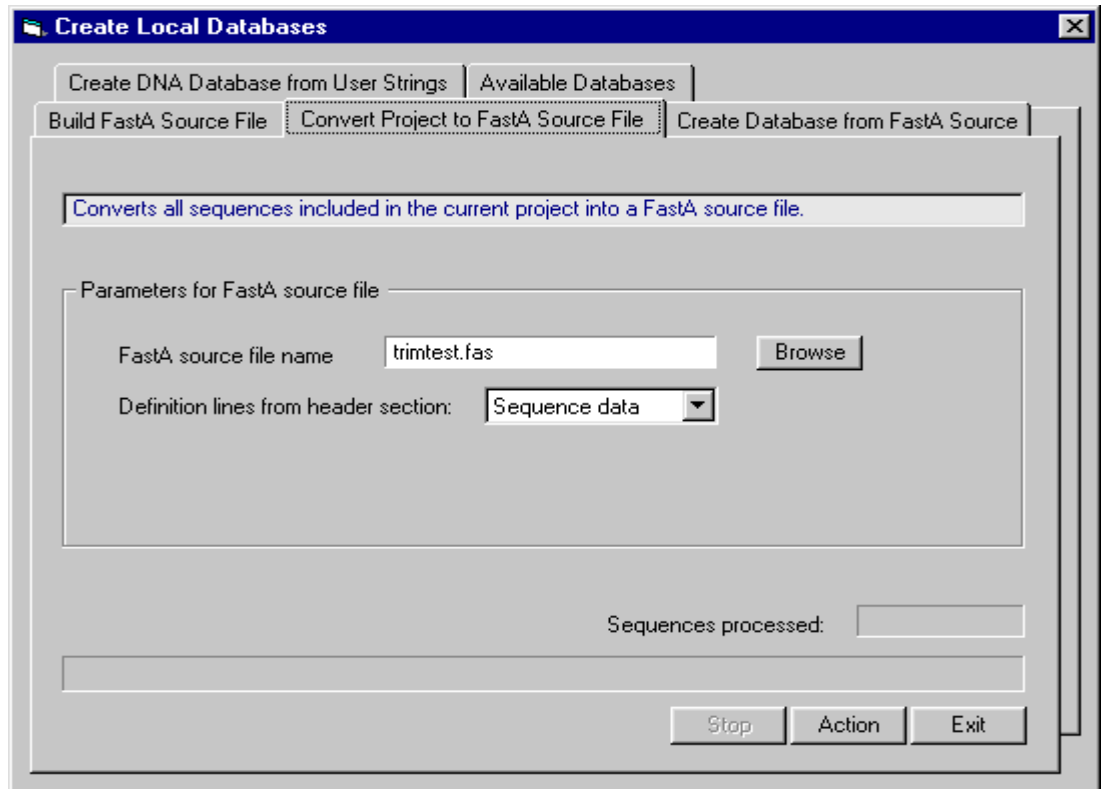
为了生成源文件，有必要收集所有的特定序列，这些序列是用户希望的包含在相同目录下的源文件中的那些序列。所有的在特殊目录下的文件都包含在数据库和不含序列的文本和程序文件中。因此在开始创建数据库之前需要仔细的审查目录的内容。在目录中的小功能 Utilities/Print Files 可用于获得一张需要被包含的文件列表。

如果序列没有被 DNAtools 注释，有必要提示哪个标题部分需要用于注释源文件和数据库



从一个 DNAtools 方案中的序列生成源文件：

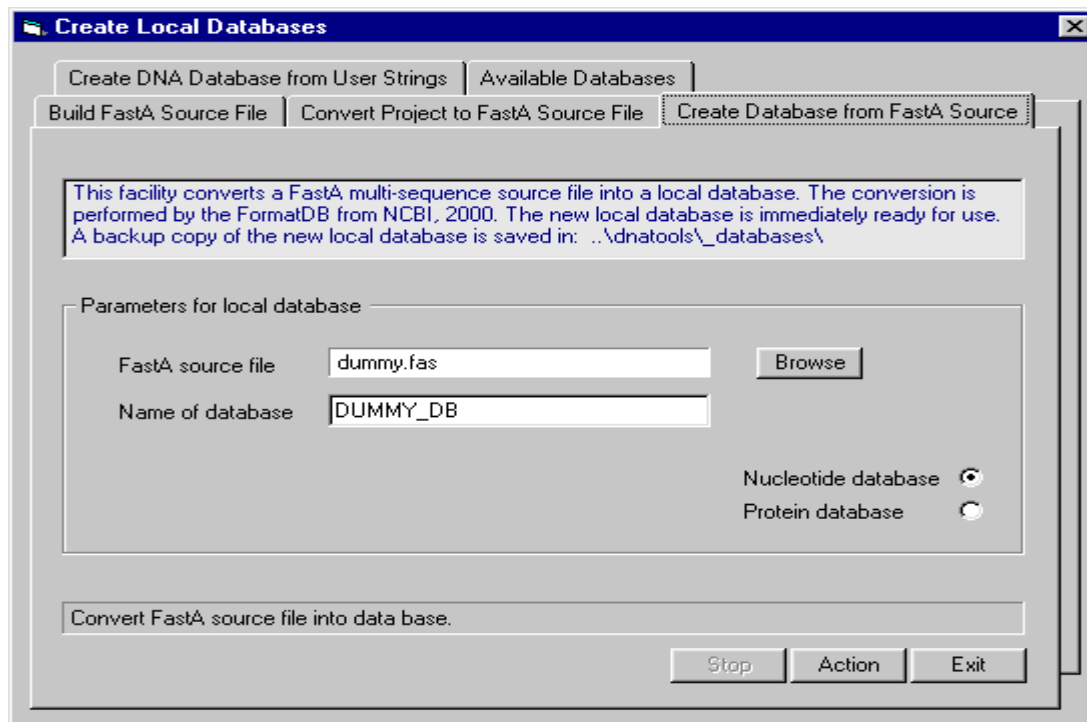
工作方式与上述相同，除了源文件是来自序列且他们的标题包含在当前方案中。



生成本地数据库（使用 formatdb，NCBI）

使用 NCBI 的 DOS 程序 formatdb 以生成可搜索的数据库。为了生成数据库，简单的选择那个你希望装入到一个可搜索数据库中的源文件，为数据库选择一个名字并选择是否这个源文件包含核苷酸还是蛋白质序列。然后点击 Action。

完成的本地数据库被创建在 Windows / Winnt 目录下的 DT5_TEMP 子目录中，同时如果希望搜索它时必须将其保留在那里。这个数据库的拷贝被保存在 \dnatools_databases 中。



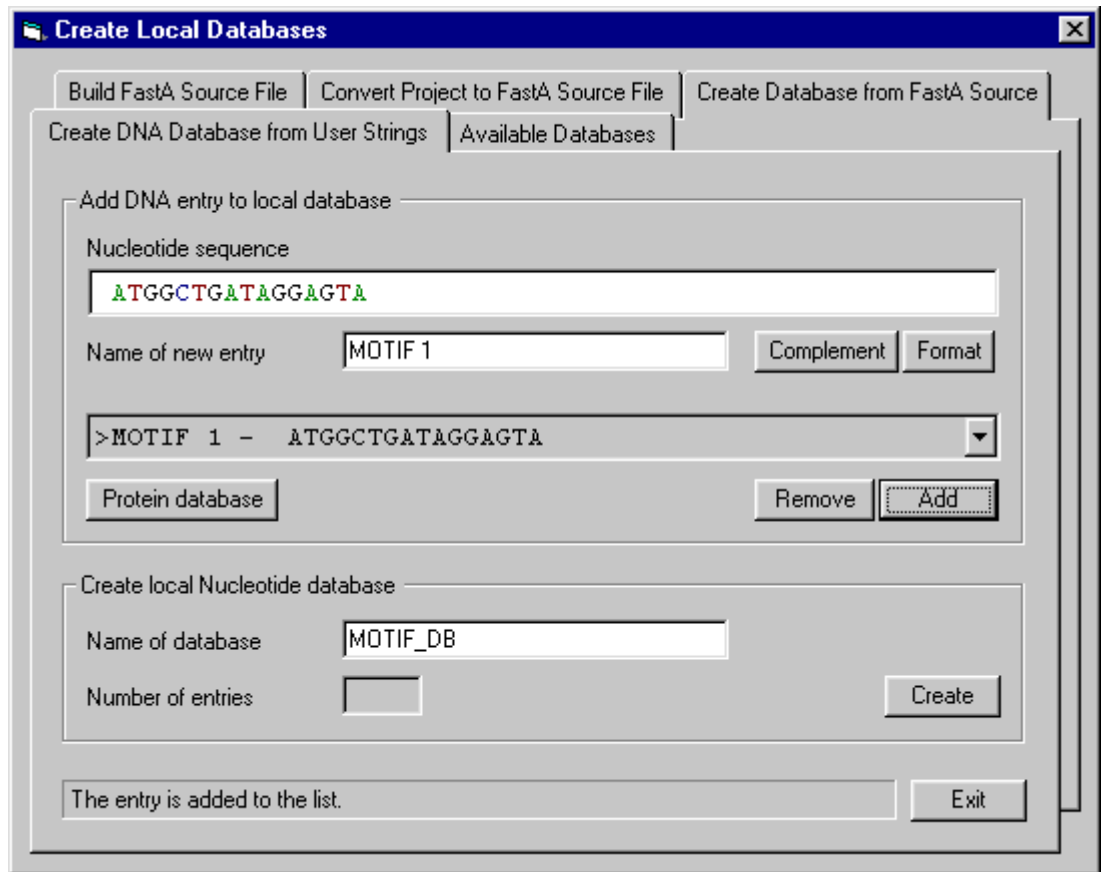
万一希望从含 DNA 序列的方案中生成一个蛋白质数据库，在建造 FastA 源文件和最终的数据库之前，使用“Utilities/Create protein files”翻译该核苷酸序列。

注释：该数据库中序列将只包含一行注释。如果输入序列是 GenBank, FastA 或 GCG 格式，注释将被自动的从初始文件中寻回。对于 GenBank 文件，使用 DESCRIPTION 行；对于 FastA，使用标准的单行注释；对于 GCG，.. 分割符号之前跟行。

对于 DNAtools 文件，列在表上的标题部分选择哪个部分用于注释。默认的是序列名，长度，审查总和和文件日期。明显的，只有真正包含信息的标题部分可被用于注释数据库。

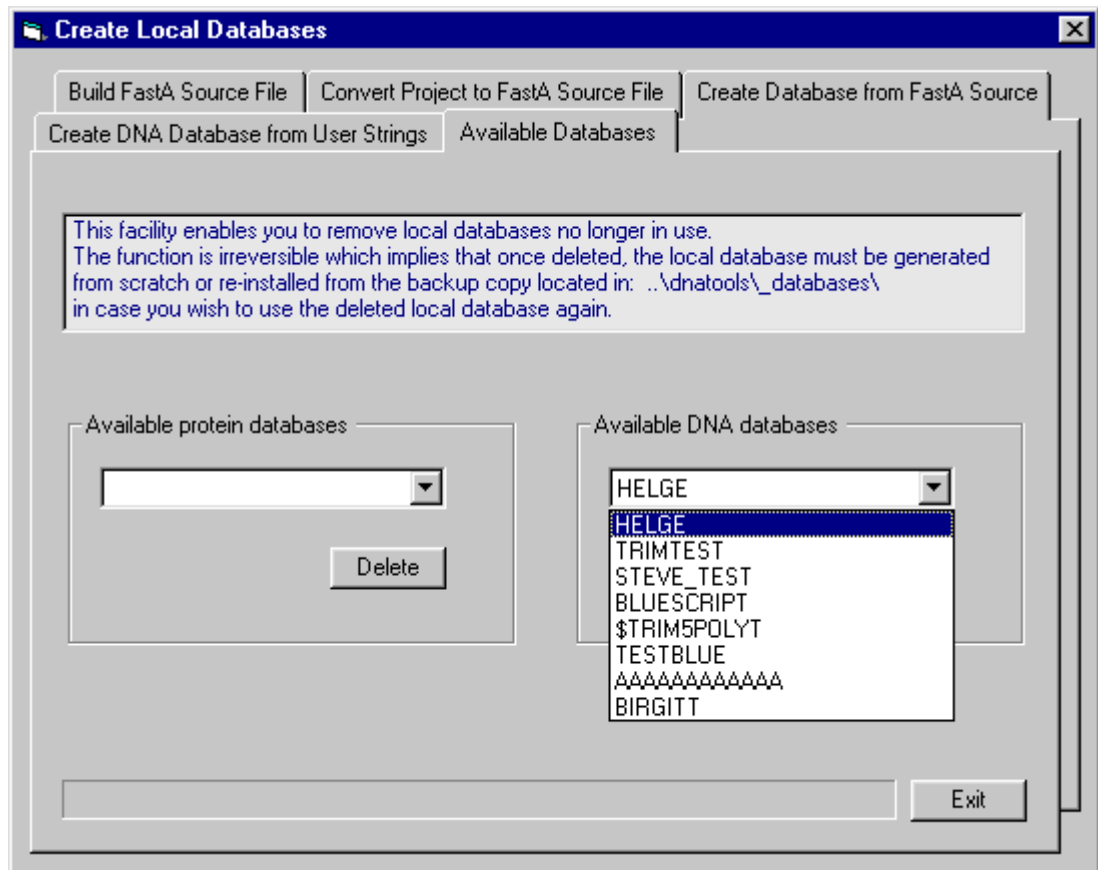
使用者列数据库：

万一用户希望从一个更小的序列基序中生成一个数据库，可以使用这个功能以输入核苷酸或蛋白质列，给每个序列命名并且最终将这些列转入到可搜索数据库中。



数据库操作：

过一些时间，本地数据库的数量可能会变得很大使得找到所希望使用的数据库变得很难。使用这个功能，可以从列表中移除数据库。注意没有“撤销”选项，因此该操作是最终的。但还是有可能恢复一个被删除的数据库，只需定位...\dnatools_databases 中 formatdb 生成的三个文件并移动他们到一般的 DOS 目录中。



Acknowledgements (感谢) 略

The blast search functions in DNATools would not have been possible without the outstanding efforts of the staff at NCBI. I would like to express my gratitude for their willingness to share their expertise with the scientific community.

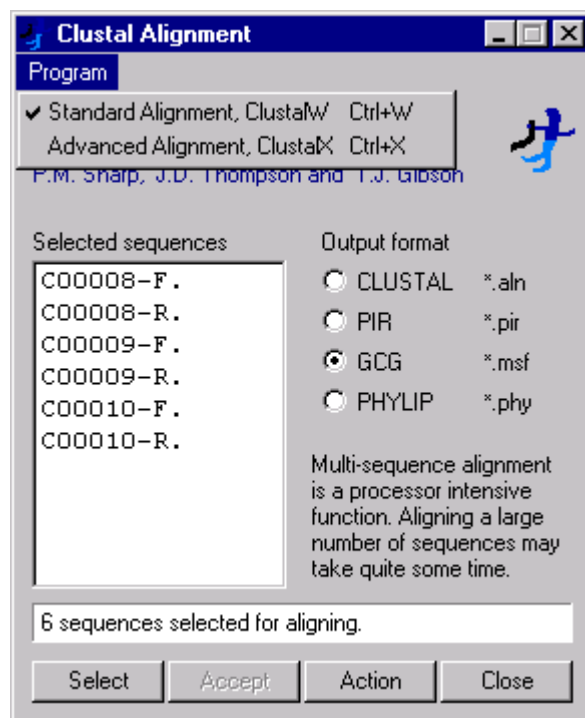
The blast programs are published by – Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389–3402.

3. 用 clustal 进行序列比对:

此项功能使用 Clustal 程序执行真正的配对。ClustalW 程序被完全整合到 DNAtools 中并且不太可能改变默认的参数，但是 ClustalX 却是卓越的视窗程序。不像 GeneDoc，可以从作者的主页直接下载它，而两个 clustal 测序只可以从 DNAtools 下载页面进行下载。

Clustal W:

执行比对按照步骤一直往下即可，按照此页底部的教程。



输出格式:

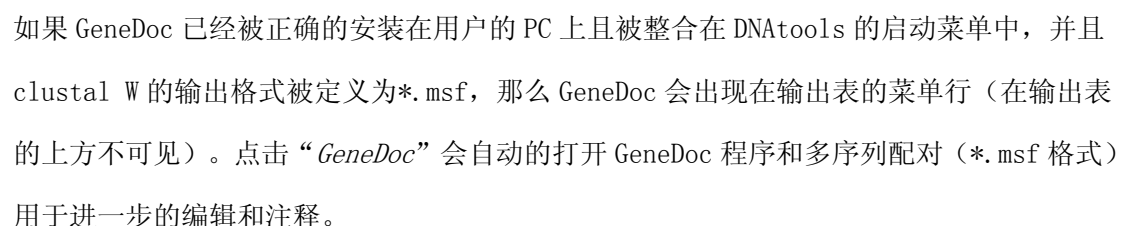
clustal 默认*.aln-默认的 clustal W 格式含方案名字和日期;

PIR, *.pir-配对的序列作为分离的序列被保存为 FastA 多序列格式;

GCG, *.msf-保存为此种格式的多序列配对可以被输入进 GeneDoc 中，一个高级配对编辑器，看如下;

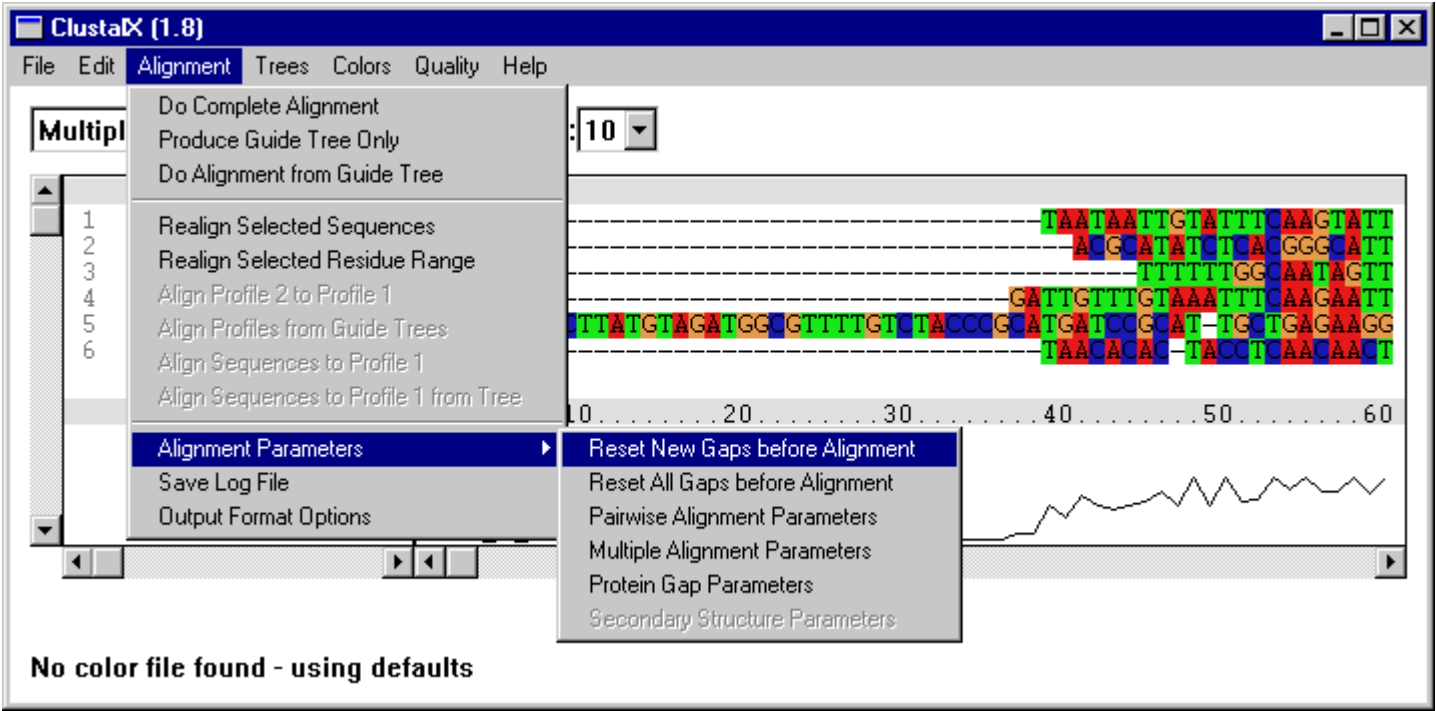
Phylip. *.phy -

来自配对的输出结果被显示在一个简单的文本编辑器中,但只含有限的选项由于注释。



Clustal X 高级配对:

Clustal X 是一卓越的视窗程序。在选择需要比对的序列之后，从 DNAtools 中选择此选项会导致序列被转载入 ClustalX 中。比对的剩余部分，包括为比对设置参数，可以在 Clustal X 程序中进行。查看下载页寻找进一步的信息。



如果用户希望进一步在 GeneDoc 中编辑 Clustal X 产生的比对，比对必须被保存为*.msf 格式。因为在 clustal X 和 DNAtools 中的 GeneDoc 之间没有直接的联系，输出的*.msf 文件必须经 GeneDoc 文件菜单被装载到 GeneDoc 中。

如何。。。。。

装载序列(DNA or protein)入一个方案中；

打开比对表，*Search/Align Sequences*；

从文件列表中选择需要比对的序列；（当按下 CTRL 键时点击左键）

接受选择；

选择输出格式；

点击 *Action* 等候;

当 *View Alignment* 命令按钮出现时, 点击它查看比对。

GeneDoc:

为了更好的利用 GeneDoc 程序, 用户必须首先下载并安装它。然后, 通过使用 DNAtools 的 *Preferences/General* 菜单在主编辑器的启动菜单中生成一个入口。这将使得 DNAtools 可以从多序列比对的结果表菜单中获得 GeneDoc。注意: 只有当输出结果是 GCG 格式时, GeneDoc 菜单选项才可以看见。

作者 (略)

Author – Multiple sequence alignment editor & Shading Utility, Version 2.5.002. , Copyright 1999 by Karl Nicholas.

从 GeneDoc 帮助文件中提取:

GeneDoc 作为一个常规的视窗程序被安装入 GeneDoc 程序组中。在启动 GeneDoc 之后, 使用选项 “file open and read in a MSF (Multiple Sequence File) file” — 在 MSF 多序列文件中打开和读取文件。GeneDoc 在 MSF 的评述部分为这些文件保存设置信息, 因此如果 GeneDoc 保存了这个文件, 它将以相同的设置被重新打开。

读取/输入数据 (帮助索引) — 用户可以输入非 MSF 文件。使用 “File/New menu” 菜单, 然后选择 “File/Import”。输入对话框允许从剪贴板、磁盘文件或手动输入。Clustal, fasta 和其他一些类型可以以这种方式被读取。

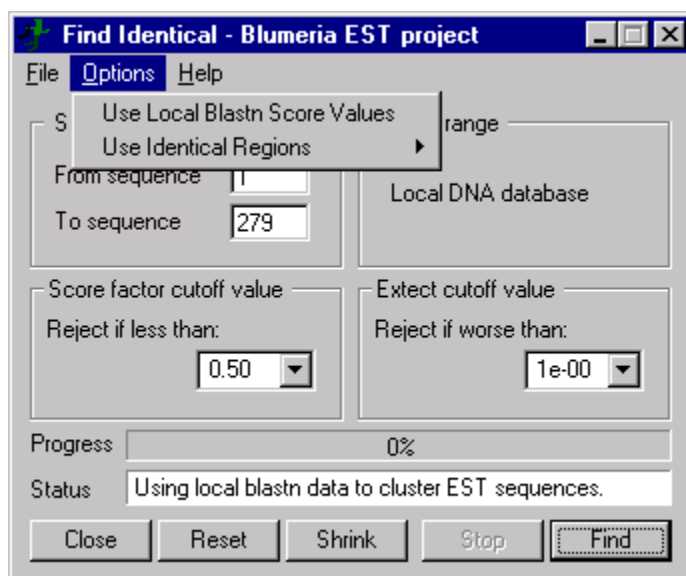
GeneDoc 网页浏览器 (帮助文档) — GeneDoc 可以在其他程序如网页浏览器或数据库程序中运行。

DDE 支持 (帮助文档) — GeneDoc 提供最小的 DDE 支持, 文件打开和打印。这将允许用户通过双击或右击 MSF 文件图标来自动的启动视窗打开文件或打印文件。

GeneDoc 可以免费获得, 无需任何保证。

4. 寻找相似性序列 blastn 数据：

此项功能使用本地 blastn 分数值评价比对，执行当前方案中所有序列的比对。在完成比对前，在序列标题中必须可以获得本地 blastn 搜索结果。例如每个序列必须包含一个本地的 blastn：在其序列标题部分。



比较包括三步：1，从所有当前方案中的序列中提取自身分数值（和自身比较时得到的分数值）；2，使用自身分数值的分割法规范每个本地 blastn 配对的分数值；3，最后，规范化后的分数值与选定的分数中止限制进行比较同时低于中止值的配对被丢弃。

期望的中止选项允许用户依据配对的期望值排除配对。在大多数情况下，中止值应该设定为 1，例如定义一个充分大的值使得所有的配对都可用于分数值评估。选项同时还包括：允许用户排除那些能产生高于某个最小值的配对的序列对。

比较结果是一张序列对列表，其本地 blastn 配对优于中止值。使用这张表以生成不同类型的报告。例子见下：

Comments 略

This function will in most cases yield the same overall clustering of EST sequences as the Find Identical Regions routine and is considerable faster. There are, however, situations where the common identical regions method will yield a more reliable clustering: Long sequences with small overlaps from the same open reading frame may be omitted when blastn clustering is used but will, if they share just a short identical region be linked by the identical region method. On the other hand, sequences from different ORFs sharing a short identical region may erroneously be linked with the identical region method but not when local blastn data are used.

如何：

使用多序列功能以生成 FastA 源文件；

使用多序列功能以转化 FastA 源文件为一个本地核苷酸数据库；

针对核苷酸数据库执行本地 blastn 搜索；

使用 “Use Local Blastn Score Values” 选项运行寻找相似序列；

生成序列或克隆报告，见下：

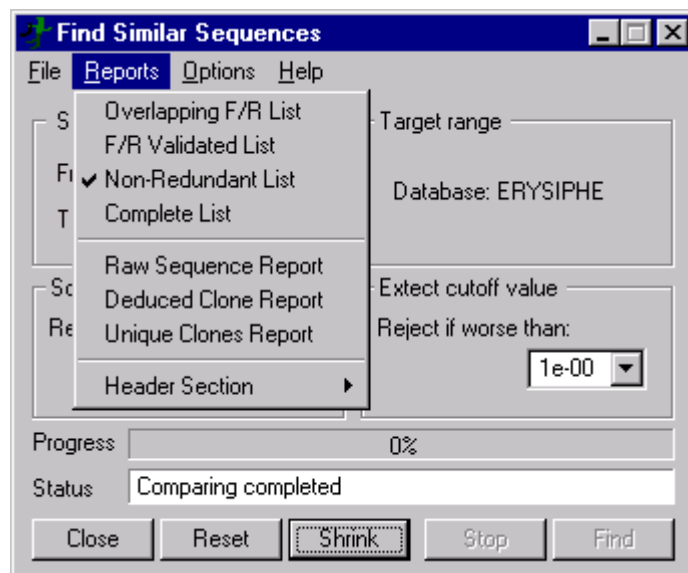
组类型：

重叠 F/R 序列—此列表包含来自相同克隆的 F 和 R 序列，其中这些克隆共享相同的区域或指定长度。对于双通过 EST 序列，提示 F 和 R 序列包含整个插入。伴随一致的序列命名，此选项可被用于显示来自同一模板的重叠序列。

R/F 验证列表-验证列表排除重叠的 F/R 序列，如上所说。由于插入太长以至于不能被前向和逆向序列或者那些特定的情况所包容如：对于一个给定的克隆，只可存在一个序列的情况。伴随一致的序列命名，此选项可被用于消除来自同一模板的重叠序列。

非多余序列列表-显示比对序列对的非多余列表。在列表中只包含一次每个比对序列对。

完整列表-按照字母显示比对序列对的完整列表。此列表包含同样的比对序列两次。暗示这些按照字母分类的列表将包含一个完整的序列列表，且这些序列和此方案中的其他序列共享一个相同的区域。



报告类型：

粗序列报告-对于输出列表中的每个比对序列对，这两个序列的名字都被与其他比对序列对的名字进行比较。万一一个比对序列对的名字和另外一个相同，将生成一个非多余名字链以产生一个粗序列报告，此报告是严格依照序列同一性。

一条链将包含所有的序列名字，且这些名字是鉴于相同配对的成员而被连接起来的。点击链的第一行，显示名字或包含于链中的序列的指定的标题行。

演绎克隆报告-演绎克隆报告是基于粗序列报告并且假定共享其文件名前 6 字符的序列指的是相同的克隆。演绎克隆报告的基本原则如下：

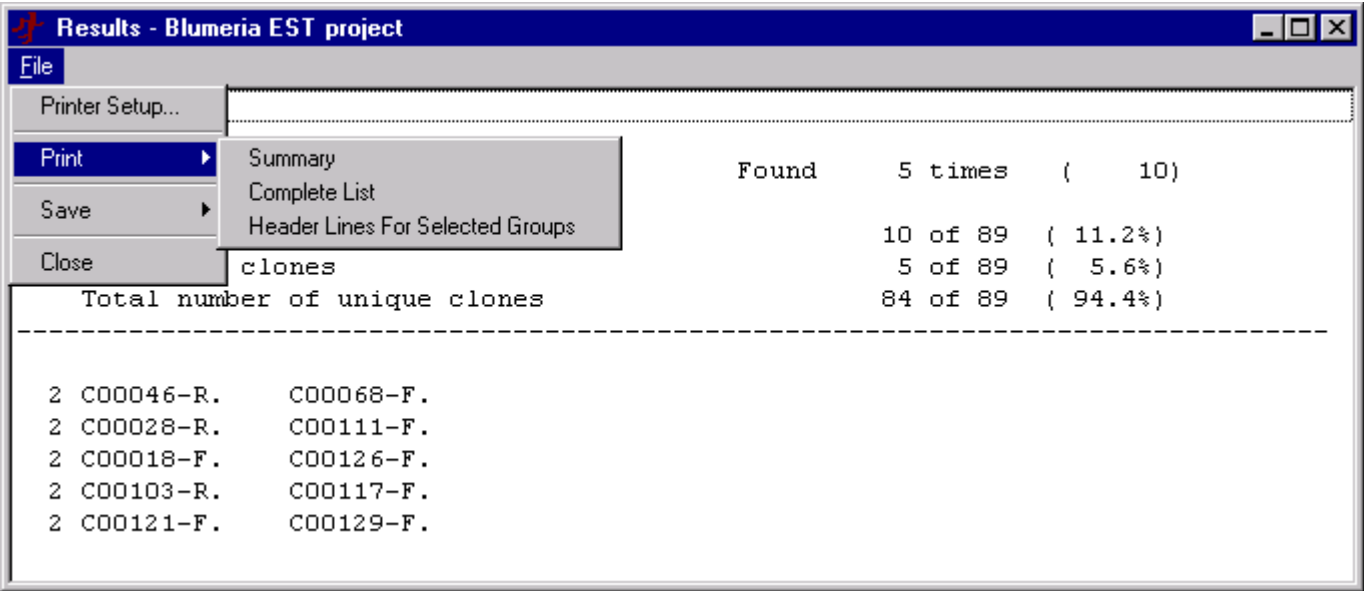
假定链 clone1-F - clone2-R 共享一个相同的序列区域；

假定链 clone2-F - clone3-R 共享一个相同的序列区域，此区域与上述的区域不同；

假定序列 clone2-R 和 clone2-F 来自同一克隆，四个序列可以被加入到一个新的链中：

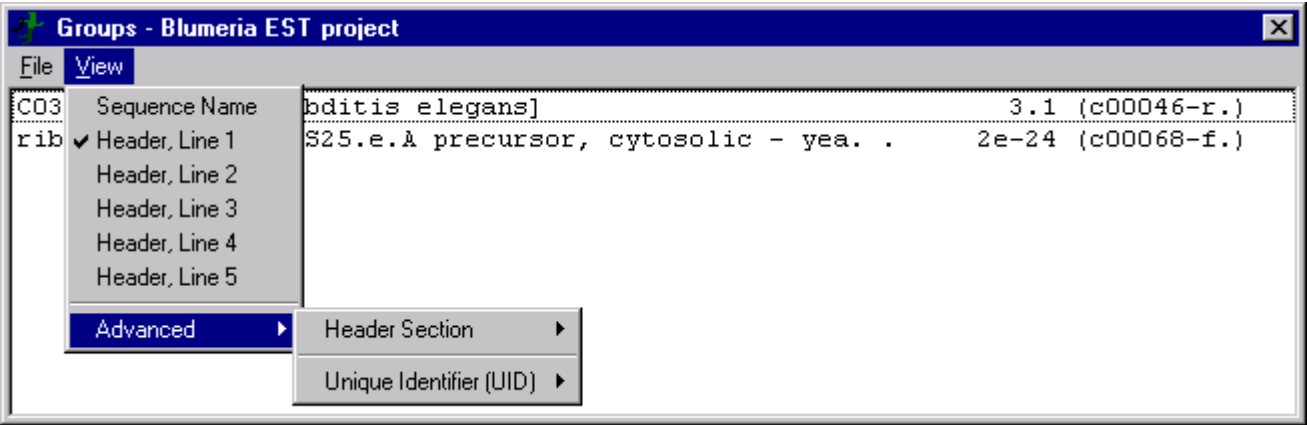
clone1-F - clone2-R - clone2-F - clone3-R；

加工那些属于新链的序列以移除来自同一克隆的完全相同的区域。通过分析序列标题和使用最小的信息标题否决序列来完成此项工作。



演绎的克隆报告可以以其在报告表中的展现形式或以一个选定的标题行列表形式被打印或保存下来，而不是以一个相同组中每个成员的克隆名形式。标题行格式与用于 View group form 中的格式相同。在保存和打印之前，首先选择格式，即打开 “Identity group form” 并选择一个 “View option”。用于打印或保存的克隆组可在报告表中被选择，即当按下 CTRL 键时点击或者拖动鼠标。当选择好组之后，点击 “File/Print/Header Line For Selected Groups 或者 File/Save/Header Line For Selected Groups.”

点击一个组的第一行，显示那个组中所有序列的注释。查看菜单，就像许多其他 DNAtools 表中的一样，允许用户定义序列标题的哪个部分用于组列表中的序列鉴定。



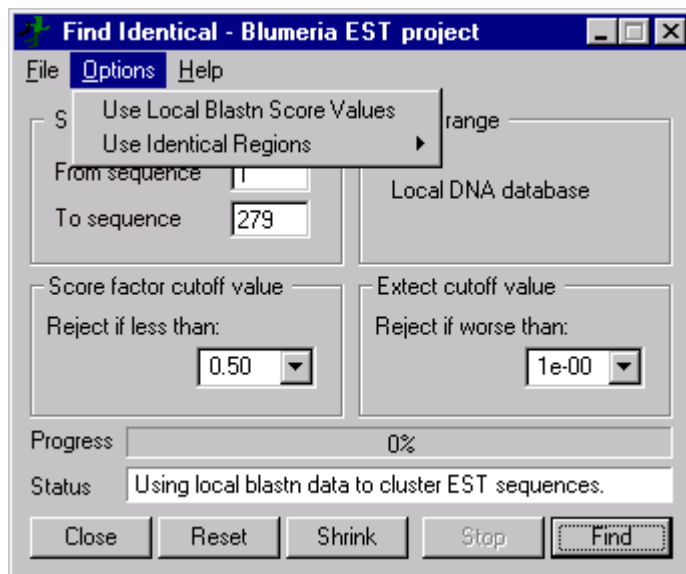
独特克隆报告—通过假定：或者直接的（就像在粗序列报告中的一样）或者通过第三个共享相同区域的序列加入方法（就像在演绎的克隆报告中的一样）的共享同一区域的两个序列都来自同一克隆，独特克隆报告是演绎克隆报告的进一步演化。独特克隆报告否决所有在链中的克隆，除了每个链的第一个克隆。这将为当前方案创建一个演绎的独特克隆列表。独特克隆报告被格式化为 microtite 板的 12*8wells 以帮助克隆阵列的生成。

关于序列名字的重要信息：

明显的，最后的两个报告类型要求文件名的前 6 个字符鉴定克隆。第 8 个字符是 F，则被定义为前向链；而 R 则被定义为逆向链。文件名的第 7 字符和文件扩展名的 3 个字符不被考虑。

5. 寻找相似性序列 相同区域：

此项功能执行当前方案中的序列比对以寻找相同的区域。



如果找到了一个匹配，匹配被延长直到两个序列不同。搜索窗口的尺寸增加以满足延展的匹配长度同时靶序列的剩余部分也被搜索。如果没有找到更多的匹配，匹配长度被报告。视窗大小被重新设置同时针对接下来的靶序列重复搜索。程序不断重复直到所有的序列都被搜索过。

当搜索完成后，搜索结果的总结和搜索范围、参数等被显示在信息框中。匹配序列列表被显示在一个独立的列表中。

在关闭信息框后，双击匹配列表中的一个项目，自动的启动序列比对功能，重复两序列比对以核实该两个序列的匹配区域。

如果搜索以一个更大的搜索视窗进行重复搜索的话，只有那些在先前搜索中找到的匹配才会被搜索。降低搜索视窗的尺寸重新定义范围为当前方案中所有序列。

如果在当前方案中的较大量序列中进行搜索且 step 长度较小，核实搜索将持续较长的时间。可点击 stop 终止搜索。若绿色条到达底端，搜索将继续。

可以延长一个已经存在的序列比对。查看“Extend identity search (*.ird file)”。

选项：

源和靶范围：搜索范围（源和靶序列）可以是当前方案中序列的任何组合。

最小匹配长度—从下拉列表框中选择一个新的值以改变搜索窗口大小（最小匹配长度），当然也可通过在文本域中输入一个值来实现。

敏感度—从下拉列表框中选择一个新的值以改变 step 长度。默认的 step 长度是视窗长度的一半但是可以被改为任何值—在 1 和视窗长度之间—列在下拉列表中。

1 个单位的 Step 长度将搜索所有可能的片断，然而等于片断长度的 step 值将以非重叠搜索窗口进行搜索。Step 长度的设置将很大的影响搜索的持续时间。

对于一个包含 2500500 碱基序列的方案，一个单位的 Step 长度搜索将花费许多小时，即使在一个快速的 PC，可能在晚上执行会更好。然而，此项工作可以在后台进行，因此可以在使用其他程序时进行比较，例如额外的 DNAtools 实例。

注意：用户可能会发现有不同的匹配数，这依赖于设置。只有 1 个单位的 step 值才会寻找所有相同的区域。一个更大的值可以减少搜索时间，但可能会错过一些相同的区域。

因为滑行窗口总是起始于序列的 5' 端并且当一个指定长度的区域不能再从序列中提取时终止，针对整个方案进行的沃森链搜索将产生一个稍微不同的结果。（相比于用互补的克里克链进行同样的搜索）。

文件菜单：

保存为—保存 R/F 核实过的，非多余列表，一个完整的列表或者克隆阵列。

打印：打印当前展示列表。

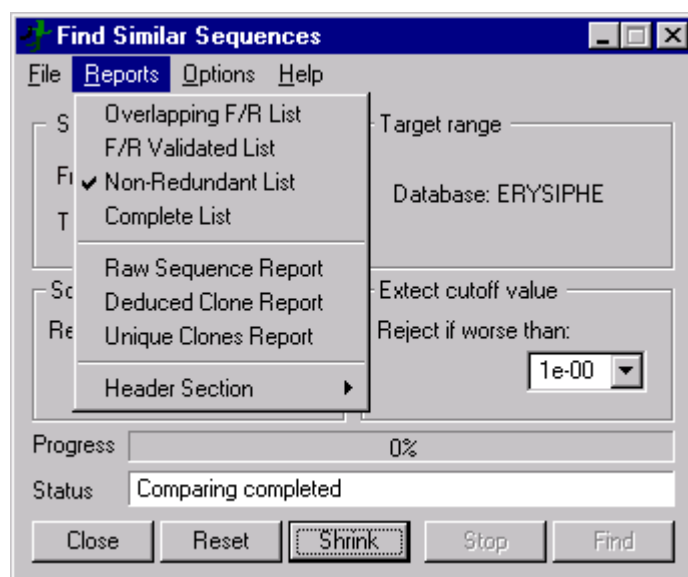
组类型：

重叠 F/R 序列—此列表包含来自相同克隆的 F 和 R 序列，其中这些克隆共享相同的区域或指定长度。对于双通过 EST 序列，提示 F 和 R 序列包含整个插入。伴随一致的序列命名，此选项可被用于显示来自同一模板的重叠序列。

R/F 验证列表-验证列表排除重叠的 F/R 序列，如上所说。由于插入太长以至于不能被前向和逆向序列或者那些特定的情况所包容如：对于一个给定的克隆，只可存在一个序列的情况。伴随一致的序列命名，此选项可被用于消除来自同一模板的重叠序列。

非多余序列列表-显示比对序列对的非多余列表。在列表中只包含一次每个比对序列对。

完整列表-按照字母显示比对序列对的完整列表。此列表包含同样的比对序列两次。暗示这些按照字母分类的列表将包含一个完整的序列列表，且这些序列和此方案中的其他序列共享一个相同的区域。



报告类型：

粗序列报告：

粗序列报告-对于输出列表中的每个比对序列对，这两个序列的名字都被与其他比对序列对的名字进行比较。万一一个比对序列对的名字和另外一个相同，将生成一个非多余名字链以产生一个粗序列报告，此报告是严格依照序列同一性。

一条链将包含所有的序列名字，且这些名字是鉴于相同配对的成员而被连接起来的。点击链的第一行，显示名字或包含于链中的序列的指定的标题行。

演绎克隆报告-演绎克隆报告是基于粗序列报告并且假定共享其文件名前 6 字符的序列指的是相同的克隆。演绎克隆报告的基本原则如下：

假定链 clone1-F - clone2-R 共享一个相同的序列区域；

假定链 clone2-F - clone3-R 共享一个相同的序列区域，此区域与上述的区域不同；

假定序列 clone2-R 和 clone2-F 来自同一克隆，四个序列可以被加入到一个新的链中：

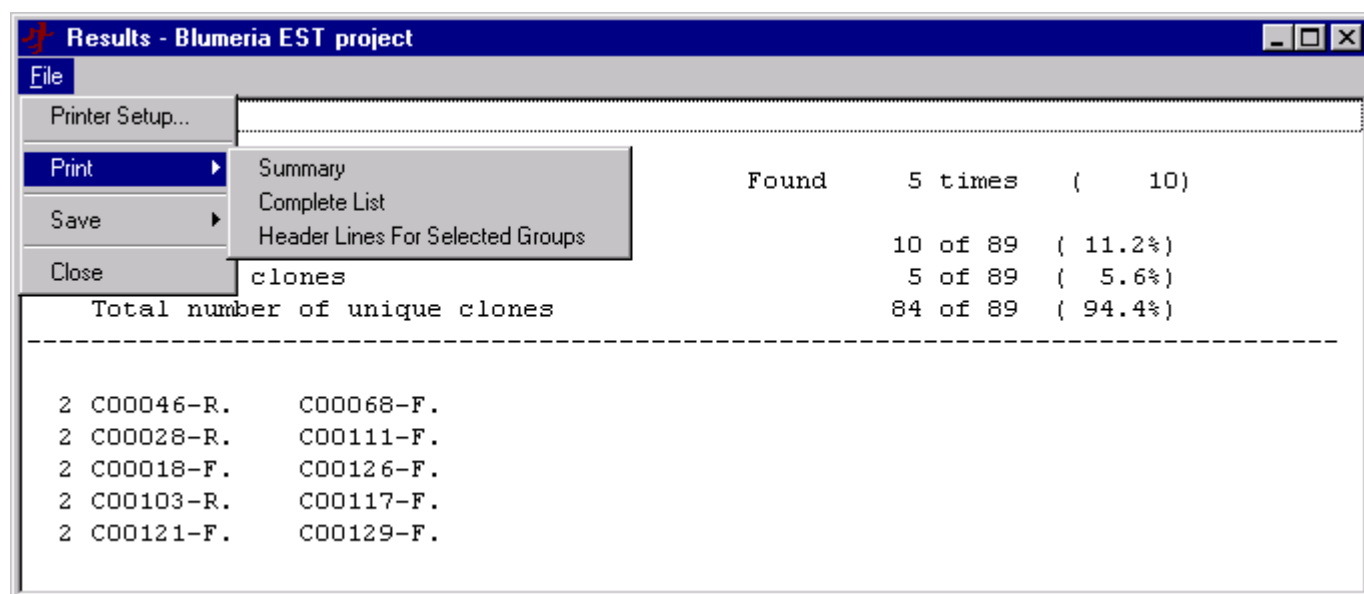
clone1-F - clone2-R - clone2-F - clone3-R；

加工那些属于新链的序列以移除来自同一克隆的完全相同的区域。通过分析序列标题和使用最小的信息标题否决序列来完成此项工作；

在演绎的克隆报告中，这两个第一链被一个单个链所取代：clone1-F - clone2-F - clone3-R（如果克隆 2-R 有最小信息标题）

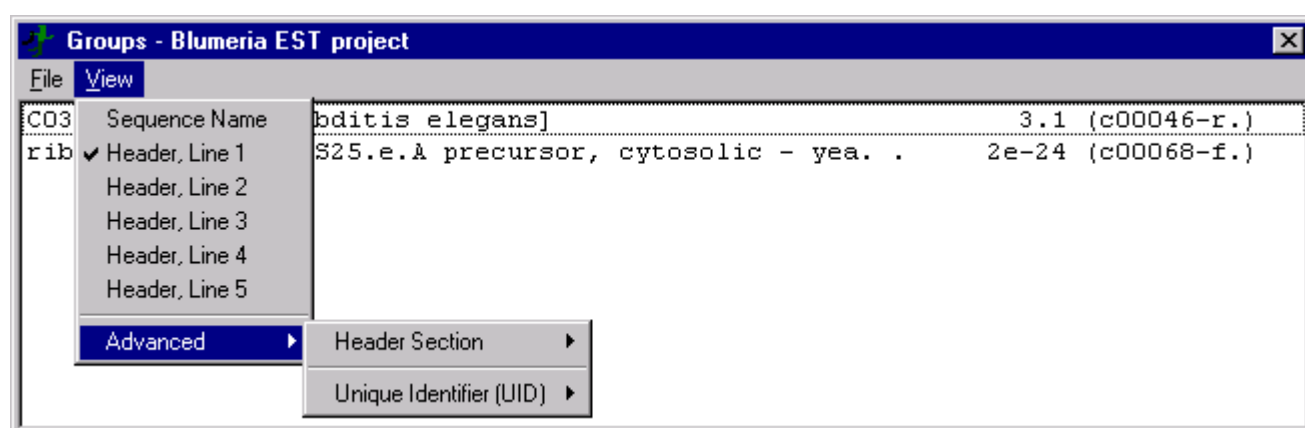
演绎的克隆报告可以以其在报告表中的展现形式或以一个选定的标题行列表形式被打印或保存下来，而不是以一个相同组中每个成员的克隆名形式。标题行格式与用于 View group form 中的格式相同。在保存和打印之前，首先选择格式，即打开“Identity group form”并选择一个“View option”。用于打印或保存的克隆组可在报告表中被选择，即当按下 CTRL 键时点击或者拖动鼠标。当选择好组之后，点击“*File/Print/Header Line For Selected Groups* 或者 *File/Save/Header Line For Selected Groups.*”

点击一个组的第一行，显示那个组中所有序列的注释。查看菜单，就像许多其他 DNAtools 表中的一样，允许用户定义序列标题的哪个部分用于组列表中的序列鉴定。



点击一个组的第一行，显示那个组中所有序列的注释。查看菜单，就像许多其他

DNAtools 表中的一样，允许用户定义序列标题的哪个部分用于组列表中的序列鉴定。



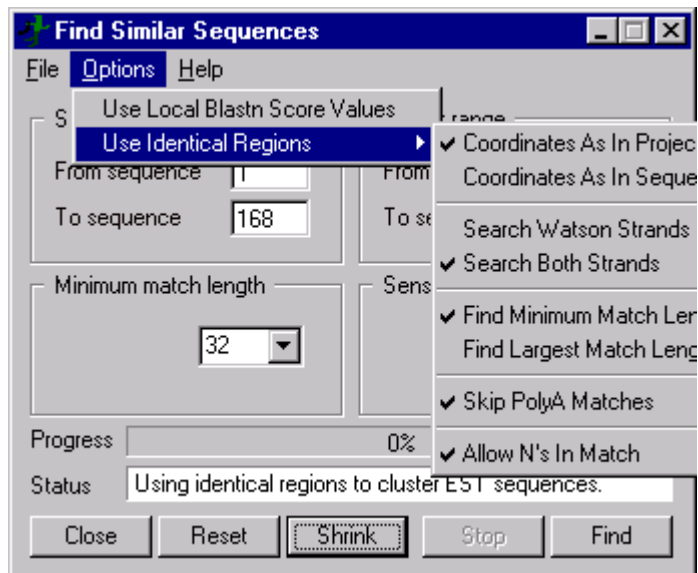
独特克隆报告—通过假定：或者直接的（就像在粗序列报告中的一样）或者通过第三个共享相同区域的序列加入方法（就像在演绎的克隆报告中的一样）的共享同一区域的两个序列都来自同一克隆，独特克隆报告是演绎克隆报告的进一步演化。独特克隆报告否决所有在链中的克隆，除了每个链的第一个克隆。这将为当前方案创建一个演绎的独特克隆列表。独特克隆报告被格式化为 microtite 板的 12*8wells 以帮助克隆阵列的生成。

明显的，最后的两个报告类型要求文件名的前 6 个字符鉴定克隆。第 8 个字符是 F，则被定义为前向链；而 R 则被定义为逆向链。文件名的第 7 字符和文件扩展名的 3 个字符不被考虑。

Coordinates:

在当前序列中—为当前方案中的序列的起始部分外加相配之物给相同的区域。如果在转换序列为其互补序列时找到了一个匹配，在结果列表中匹配将被标记为 C。

在匹配序列中—为源序列（在这些序列中找到了匹配）的起始部分外加相配之物给相同的区域。为了在方案中的序列中定位匹配，在克里克链上找到的匹配（被 C 所标记的）必须先被转换为其互补序列。



链:

沃森链—用源序列的沃森链进行搜索。

两条链—搜索源序列的两条链。

区域:

找到最大相同区域—此选项最耗时间，除非用户真的期望看看这些最长的相同区域，否则还是不要使用它。

只找视窗长度一紧紧寻找但是并不延展最初的相同区域。选项是默认的并且是优先使用的以生成序列相同区列表。双击列表中的一个项目激活比较功能，这将产生一个完整的相同区域列表。

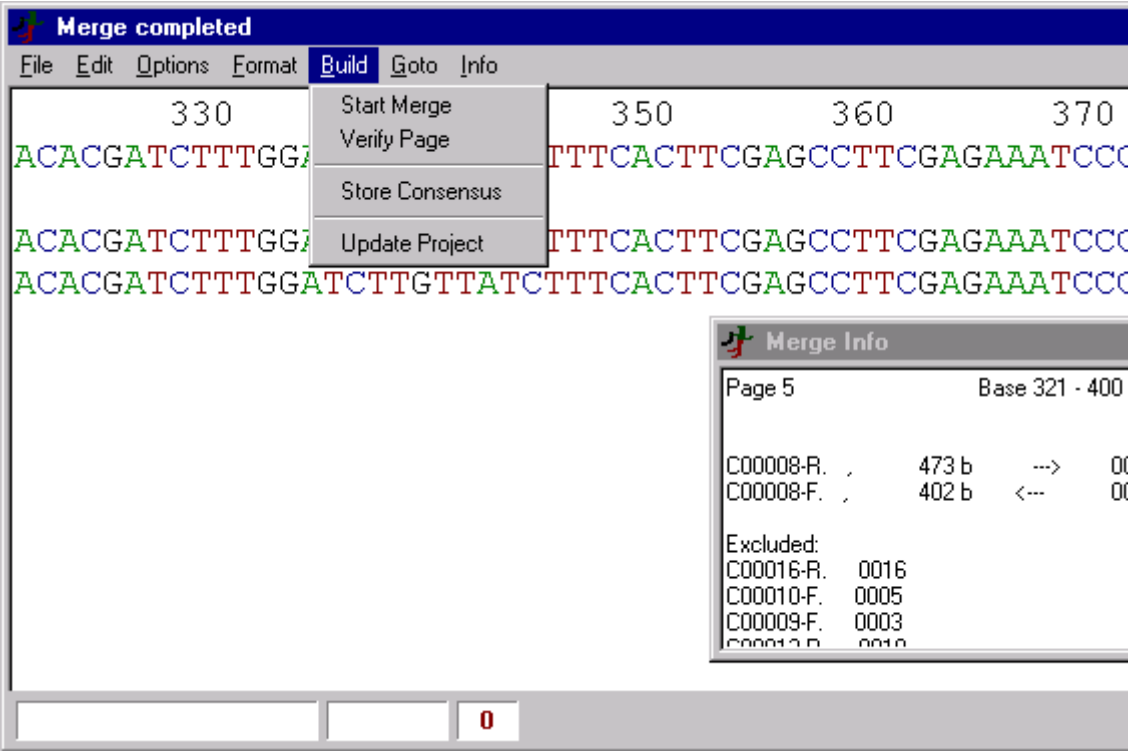
关于序列名字的重要信息：

明显的，最后的两个报告类型要求文件名的前 6 个字符鉴定克隆。第 8 个字符是 F，则被定义为前向链；而 R 则被定义为逆向链。文件名的第 7 字符和文件扩展名的 3 个字符不被考虑。

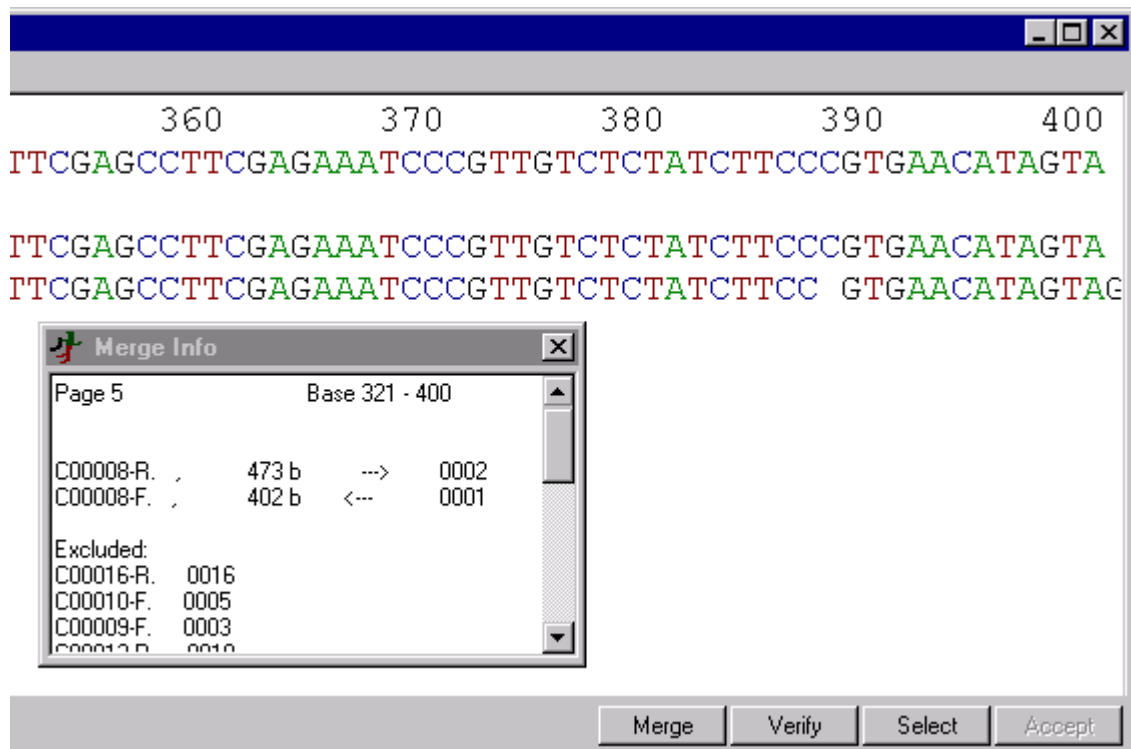
6. 连接编辑：

此项功能辅助用户建造和编辑重叠核苷酸序列的合并部分。

连接编辑器菜单：



命令按钮用于选择，核实和合并序列：



队列寻求子序列所有匹配的最大相同区域。然后使用这些信息展示合并的，排列的相同区域。在队列加工过程中，并没有考虑队列区域上游和下游的序列，提示：只有在进行以下操作即：“在合并子序列之前，编辑这些亚序列以去除低质量区域”时这项功能方才有用。同时要求亚序列属于同一连接，如来自同样的基因/开放读码框/核苷酸片断。

Page up 和 page down 移动序列队列到队列的下或前一页；

按下 CTRL 显示一个小的文本域用于输入序列序号，按下回车键进入到选择的基数；

在合并中使用箭头键四周移动，序列名、长度和当前鼠标位点的基数被显示在合并下的两个文本域中；

所有包含在当前页中的序列和其在合并的定位被显示在不同的表中，其中使用从合并中排除出来的序列的名字和箭头提示这些定位；

在合并编辑中那些被升级后的编辑序列必须被保存为一个常规的方案，如果用户希望保留这些改变的话；

合并编辑器接受至少 200,000 碱基的序列 (on a Pentium 266 MHz, 98 Mb RAM).

如何。。。。。

装载那些用户希望合并进入 DNA 序列方案中的序列;

点击选择以显示文件列表;

高亮显示将被合并的序列 (在点击时按下 CTRL 键);

点击 Accept 以包含在合并中的高亮序列;

点击 Merge 执行;

点击 Update project 以使用进入合并中的变化升级方案;

点击 *Store Consensus* 将同意附加到当前方案中;

从主菜单中选择每行的基数;

点击 *File/Print Merge*. 打印合并。

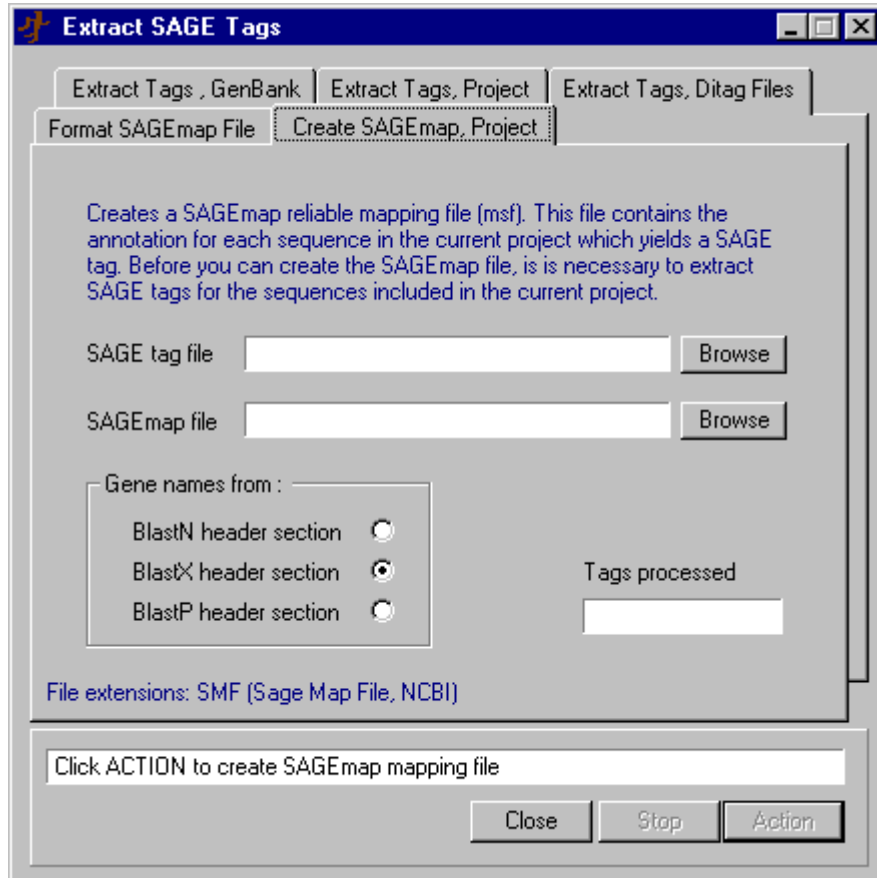
评述: 略

7. SAGE 可靠的映象文件:

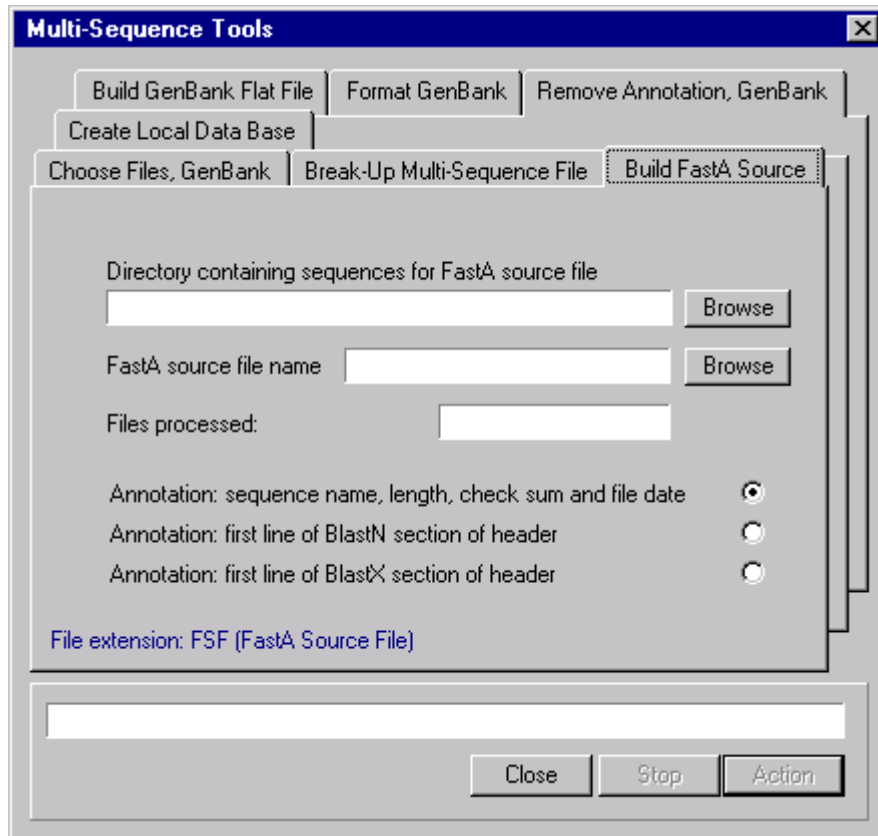
为了利用注释选项, 用户需要生成/下载一个 SAGEmap (*. smf) 并且使用这个文件中的数据以鉴定相应于 SAGE 标记的基因。

基本上, 一个位图文件是一标号定义的 ASCII 文件 (包含标记序列), 独特的基因标记子/克隆名字和注释行/基因名。

位图文件可以从 NCBI 下载, 人类的、小鼠的、大鼠的和 *S. cerevisiae*. 都可以。如果用户对另一个有机体进行研究, 可以在注释的 EST 库基础上构建一个位图文件或者从一个 FastA 多序列文件中构建一个位图文件。DNAtools 包含多项功能可以用于生成来自两种数据类型 SAGEmap。



如果不能获得 FastA 多序列文件，可以用多序列功能生成一个。



Example of a SAGemap reliable mapping file

```

AAAAAAAAA C00196-R heat shock protein 70 [Trichophyton rubrum] 2e-16
AAAAAAAAA C00224-F protein associated with DNA helicase/prim. . 6.0
AAAAAAAAA C00280-R hypothetical protein Rv2052c [Mycobacte. . 0.37
AAAAAAAAA C00822-M HYPOTHETICAL 24.1 KD PROTEIN C17A5.08 IN CH. .
9e-19

AAAAAAAAA C01407-R No description list for sequence C01407-R.
AAAAAAAAA C0A12-1R mucin, tracheobronchial - dog
>gi|402558|emb|CAA4891. . 8.5

AAAAAAAAA D00131-F No description list for sequence D00131-F.
AAAAAAAAA D00369-F 64aa long hypothetical protein [Aerop. . 0.008
AAAAAAAAA D00428-R No description list for sequence D00428-R.
AAAAAAAAA D00470-M No description list for sequence D00470-M.

```

AAAAAAAAAA D00581-F HEAT SHOCK PROTEIN HSP1 (65 KD IGE-BINDING . .
 6e-44
 AAAAAAAAAA D00599-M No description list for sequence D00599-M.
 AAAAAAAAAA D00620-F TYPE II DNA MODIFICATION ENZYME (METHYLTRA. .
 0.36
 AAAAAAAAAA D00762-M HYPOTHETICAL 37.2 KD PROTEIN IN ALG9-RAP1 I. .
 6e-04
 AAAAAAAAAA D00818-F No description list for sequence D00818-F.
 AAAAAAAAAA D00837-F PUTATIVE GLUCOSYLTRANSFERASE C08H9.3 >gi|38. .
 6.3
 AAAAAAAAAA D00940-M endonuclease [Magnaporthe grisea] 5e-53
 AAAAAAAAAA D01107-M A2-5a orf23; hypothetical protein [Ba. . 1.7
 AAAAAAAAAA D01268-F GTP-binding protein ypt5 - fission yeast
 (Schizosacc. .2e-12
 AAAAAAAAAA D01294-M glycoprotein [Vesicular stomatitis virus] 9.5
 AAAAATCTTG D00950-M LONG-CHAIN-FATTY-ACID--COA LIGASE 3 (LONG-C. .
 7e-10